International Conference on Computer Sciences

IHCONCS 2026

PROCEEDINGS BOOK

TAM METİN KİTABI



NEU PRESS

Publisher: NECMETTIN ERBAKAN UNIVERSITY PRESS (NEU PRESS)
Yaka Mah. Yeni Meram Cad. Kasım Halife Sok. B Blok No: 11 Meram / KONYA / TÜRKİYE
+90 332 221 0 575 www. neupress. org info@neupress.org
Certificate Number: 49846
ROR ID: https://ror.org/013s3zh21
Crossref Funder ID: 501100016981
Executive Editor:
Dr. Öğr. Üyesi/Asst. Prof. Dr. Fatih KALECİ, fkaleci@neupress.org

Publication Number: 544

Book Title: IHCONCS 2026 PROCEEDINGS BOOK - INTERNATIONAL CONFERENCE ON COMPUTER SCIENCES

E-ISBN: 978-625-5002-31-0

Editors
Dr. Adem TEKEREK
Dr. Jana ŽILJAK GRŠIĆ
Dr. Diana BRATIĆ
Dr. Erol KURT
Dr. Mehmet TEKEREK

Design
Enes Malik YALÇIN, enesyalcin@neupress.org
Printing Officers
Muhammed Sami TEKİN, muhammedsami@neupress.org
Nurullah DEMİR, nurullah@neupress.org
Printing
Necmettin Erbakan University Press (NEU PRESS)
Yaka Mah. Yeni Meram Cad. Kasım Halife Sok. B Blok No: 11 Meram/KONYA/TÜRKİYE
0332 221 0 575 - www. neupress.org
Certificate No: 48888
Publication Type: Book-Academic Publications
Publication Date: October 2026
Place of Publication: Konya, Türkiye
Language: English
Dimensions: 21x29. 7 cm
Edition: 1
Page Count: XXIV+1525
Category: Social, Humanities, and Administrative Sciences
SDG Category: 9 - Industry, Innovation, and Infrastructure
Keywords: Computer Science, Artificial Intelligence, Machine Learning, Cybersecurity.

OA License: CC BY 4. 0
Access: neupress. org
Copyright 2026

**This book is published under CC BY license, which means that you can copy, redistribute, remix, transform, and build upon the content for any purpose,* even commercially, as long as you give appropriate credit, provide link to the license, indicate if changes were made, and do not impose additional terms or conditions on others that prohibit them from exercising the rights granted by that license, including any effective technological measures.

Conference Committee

General Chairman
Dr. Adem Tekerek
General Chair | Gazi University, Ankara, Türkiye

Co-Chairmen
Dr. Jana Žiljak Gršić
Organizing Committee Member | Zagreb University of Applied Sciences (TVZ), Zagreb, Croatia

Dr. Diana Bratić
Organizing Committee Member | University of Zagreb, Zagreb, Croatia

Organizing Committee
Dr. Erol Kurt
Organizing Committee Member | Gazi University, Ankara, Türkiye

Dr. Mehmet Tekerek
Organizing Committee Member | Ankara University, Ankara, Türkiye

International Technical Committee

- **Dr. Abzetdin Adamov**, ADA University, Azerbaijan
- **Dr. Ahmet Artu Yıldırım**, Advanced Micro Devices, United States
- **Dr. Ali Safaa Sadiq**, Nottingham Trent University, United Kingdom
- **Dr. Amit Bhagat Maulana**, Azad National Institute of Technology, Bhopal, India
- **Dr. Amir Pourabdollah**, Nottingham Trent University, United Kingdom
- **Dr. Ayhan Erdem** , Gazi University, Türkiye
- **Dr. Bünyamin Ciylan** , Gazi University, Türkiye
- **Dr. Burak Küçükgöz**, Newcastle University, United Kingdom
- **Dr. Cemal Gemci**, Cyberverse, Türkiye
- **Dr. Cemal Koçak** , Gazi University, Türkiye
- **Dr. Daan Van Den Berg** , Vrije Universiteit Amsterdam, Netherlands
- **Dr. David Joseph Brown**, Nottingham Trent University, United Kingdom
- **Dr. Davor Cafuta**, Zagreb University of Applied Sciences (TVZ), Croatia
- **Dr. Dharmendra Singh Rajput**, Vellore Institute of Technology, Vellore, India
- **Dr. Eda Akman Aydın** , Gazi University, Türkiye
- **Dr. Elviz Ismayilov**, Azerbaijan State Oil and Industry University, Azerbaijan
- **Dr. Fecir Duran**, Gazi University, Ankara, Türkiye
- **Dr. Fakhranda Alimardanova**, Khazar University, Azerbaijan
- **Dr. Hasan Tınmaz**, Woosong University, South Korea
- **Dr. Hüseyin Çakır** , Gazi University, Türkiye
- **Dr. Hüseyin Polat** , Gazi University, Ankara, Türkiye
- **Dr. Ivica Dodig**, Zagreb University of Applied Sciences (TVZ), Croatia
- **Dr. İsmail Şahin**, Gazi University, Türkiye
- **Dr. Kayode Owa**, Nottingham Trent University, United Kingdom
- **Dr. Kamal Eddine MELKEMI**, University of Batna2, Algeria
- **Dr. Krešimir Rakić**, University of Mostar, Bosnia and Herzegovina
- **Dr. Lidija Tepeš Golubić**, Zagreb University of Applied Sciences (TVZ), Croatia

- **Dr. Lidia Yadira Pérez-Aguilar**, Autonomous University of Sinaloa, Mexico
- **Dr. Mădălin-Dorin Pop** , Politehnica University of Timișoara, Romania
- **Dr. Magdalena Palacz**, Silesian University of Technology, Poland
- **Dr. Mehmet Önder Efe** , Hacettepe University, Türkiye
- **Dr. Mehmet Gök** , İstiklal University, Türkiye
- **Dr. Mehmet Şimşek**, Sinop University, Türkiye
- **Dr. Marko Maričević**, University of Zagreb, Faculty of Graphic Arts, Croatia
- **Dr. M. Kamrul Hasan** , Universiti Kebangsaan Malaysia, Malaysia
- **Dr. M. Ali Akcayol**, Gazi University, Ankara, Türkiye
- **Dr. M. Mutlu Yapıcı** , Ankara University, Türkiye
- **Dr. M. Sabarimalai Manikandan**, Indian Institute of Technology Bhubaneswar, India
- **Dr. M. Shamim Kaiser**, Institute of Information Technology, Bangladesh
- **Dr. M. Mostafizur Rahman**, American International University, Bangladesh
- **Dr. Mufti Mahmud**, King Fahd University of Petroleum and Minerals, Saudi Arabia
- **Dr. M. Arifur Rahman**, Nottingham Trent University, United Kingdom
- **Dr. Murat Aydos** , Hacettepe University, Türkiye
- **Dr. Murat Dörterler** , Gazi University, Türkiye
- **Dr. Necaattin Barışçı** , Gazi University, Ankara, Türkiye
- **Dr. Nursel Yalçın** , Gazi University, Türkiye
- **Dr. Ömer Faruk Bay** , Gazi University, Türkiye
- **Dr. Pedro Machado**, Nottingham Trent University, United Kingdom
- **Dr. Pratik Vyas**, Nottingham Trent University, United Kingdom
- **Dr. Pipit Utami**, Central University of Jharkhand, Indonesia
- **Dr. Rajesh Kaluri** , Vellore Institute of Technology, Vellore, India
- **Dr. Recep Başarıcı** , İstiklal University, Türkiye
- **Dr. Shamim Al Mamun**, Jahangirnagar University, Bangladesh
- **Dr. Suat Özdemir** , Hacettepe University, Türkiye
- **Dr. Sinan Toklu** , Gazi University, Türkiye
- **Dr. Saadin Oyucu** , Gazi University, Türkiye
- **Dr. Stjepan Tvorić**, Zagreb University of Applied Sciences (TVZ), Croatia
- **Dr. Tibor Skala**, University of Zagreb, Faculty of Graphic Arts, Croatia
- **Dr. Thippa Reddy Gadekallu**, Zhongda Group, Jiaxing, China
- **Dr. Yılmaz Vural**, University of California, United States
- **Dr. Yusuf Sönmez** , Gazi University, Ankara, Türkiye
- **Dr. Zafer Ayaz** , Gazi University, Türkiye
- **Dr. Željko Kovačević**, Zagreb University of Applied Sciences (TVZ), Croatia

Secretariat Committee

- **Semiha Güngör , Hacı Bayram Veli University, Ankara, Türkiye**
- **Büşra Duygu Çelik , Gazi University, Ankara, Türkiye**
- **Öznur Şifa Akçam , Karatekin University, Çankırı, Türkiye**

TABLE OF CONTENTS

1. Innovative yet Low-Resource Uses of AI in Teaching: Pedagogical Boundaries and Ethical Considerations for Teachers 1
2. A Methodological Framework for Preparing Future Computer Science Teachers for Sustainable Professional Activity in the Artificial Intelligence Context 10
3. Gender Pay Gaps in Software Engineering: A Literature Review of Attrition, Measurement, and Structural Inequality 21
4. AI-Driven Dynamic Trust Scoring over Self-Sovereign Identity Systems 31
5. Digital twin-based residual analysis for leak detection and localization in high-pressure gas pipelines 40
6. Evaluating Quantum Computing Use Cases in Banking: A Sequential Model-Risk Admissibility Framework 49
7. A New Topological Index-Based Solution Approach for the Traveling Purchaser Problem 59
8. Machine Learning in Agile Project Management: A Bibliometric Analysis 68
9. Real-Time Helmet Detection System with Monitoring and Alert for Restricted Areas 80
10. Efficient Quantum Coherence/Decoherence Emulation Using Parallel-Recursive Computation on FPGAs and Algorithmic Skeletons 90
11. Mining Sequential Rules from Large-Scale Sequence Databases Using an FPTree-Based Method 102
12. Contemplating the Effects of NP-Complete Reductions on Problem Instance Hardness 114
13. Explainable Commit-Level Vulnerability Intelligence in DevSecOps Pipelines: A Systematic Review and Experimental Analysis 131
14. The Illusion of Cultural Competence: Evaluating ChatGPT-Generated Kazakh Proverbs 143
15. A QUBO-Based Reformulation and Quantum Solution Framework for Nonlinear Constrained Optimization Problems 153
16. Long Distance Environmental and Security System 164
17. Comparative Evaluation of Heuristic Search Algorithms for Reward-Constrained Museum Visitor Routing on a Real-World Graph 174
18. Cybersecurity Awareness and Threat Perception Constructs 185
19. Model Predictive Control-Based Trajectory Tracking and Dynamic Obstacle Avoidance for an Autonomous Vehicle in Robot Operating System 2 195
20. From Judgment to Benchmark: An AI-Assisted Wedge Classification Framework for Commercial Loan Spread Governance 204
21. Evolution in Remote Sensing Change Detection: Methods, Challenges, and Next AI Perspectives 220
22. Multi-Objective Hybrid Swarm Optimization for Energy-Delay Trade-off in SDN-enabled MEC 231
23. Enhancing Small Object Detection in Mammography Using a Patch-Based YOLO Framework 242
24. From Regulatory Compliance to Digital Trust: Integrating NIS2 and ISO/IEC 27701 In Transitional Cybersecurity Governance - The Case of Bosnia and Herzegovina 255
25. Reward System For Student Social Activity: Lntu Case Study 269
26. Adaptive Lossless Compression for Event-Based Cameras 275
27. Artificial Intelligence in Online Gambling and Betting Systems: A Knowledge Representation-Based Use Case 284
28. Spatiotemporal Pattern Detection Of Precipitation Using Machine Learning And Remote Sensing Data 293
29. Beyond Aggregate Metrics: Population-Weighted CVaR Auditing for Hidden Performance Failures Along Continuous Attributes 303

30. A Dual-Explainer GRU–TCN for Cyber-Physical Attack Detection in IoT-Based Energy Infrastructure 313
31. Boundary-Contact Modeling of Tumor Growth with Neural Surrogate Acceleration for Embedded Biomedical Systems 322
32. Decision Tree-Based Data Completion in Structured Lexicographic Resources 329
33. A Comparative Analysis of Vision Transformer Models for Multimodal Emotion Recognition 341
34. Intelligent Cyber Threat Detection at Scale: A Comparative Study of Machine Learning and Deep Learning Approaches for Network Intrusion Detection on CICIDS2017 351
35. Hybrid Deep Learning Model and Zonal Clustering for RNA 3D Structure Prediction 362
36. Explainable Multi-Type Loneliness Detection from Reddit Using Classical ML, Neural Networks, and Transformer Models with LIME and SHAP Analysis 372
37. Continuously Adaptive Digital Twins for Sim-to-Real Learning in Agro Food Systems 382
38. Water Quality Monitoring and Pollution Detection Using Edge Internet of Things-Based Hybrid Artificial Intelligence Models: A Systematic Literature Review 395
39. A Practical Evaluation of CNN Architectures for Histopathological Lung Cancer Classification 412
40. Edge-Based Digital Twin for Real-Time Energy Monitoring in Food Manufacturing: A Case Study of a Bakery Enterprise 419
41. Machine Learning Models for Monitoring and Analysis of Mobile Object Movements at a Food Industry Enterprise 425
42. Explainable Mammography Classification via Attention-Guided SE-EfficientNet-B3 and Grad-CAM Diagnostic Interpretation 436
43. Artificial Intelligence in Business: Opportunities and Risks 445
44. An Analysis of the Ministry of Health of the Republic of Türkiye from a Web Accessibility (Barrier-Free Design) Perspective 457
45. Harnessing Data Science and Machine Learning for Financial Fraud Detection Using the Central Bank of Libya Dataset 468
46. Wavelet-Based Machine Learning for Real-Time Burst Detection on Embedded Edge Devices 479
47. Role-Based Phishing to Ransomware Readiness for Clinical Staff 489
48. Explainable Machine Learning for Heart Disease Risk Prediction: Comparative Evaluation of ML Algorithms, Feature Importance Stability, and Clinical Interpretability 500
49. Preliminary Development of an Aircraft Structural Damage Classification System Using Deep Learning 510
50. From Neo4j to Redis: Storing Graph Data and Comparing Graph Algorithm Performance 519
51. Open Cloud Voice Service Architecture 528
52. Evaluating the Effectiveness of the AIM-5 Model for AI Teaching Self-Efficacy and AI Integration Readiness 538
53. Generative AI through the Student Eye: perspectives, attitudes, and usage 548
54. Analysis Of Existing Approaches In The Field Of Cryptographic Protocols In Uavs 559
55. SHRUMAC: An IoT-Smart Hygiene, Refuse Monitoring and Alert Controller System for Real-Time Institutional Campus Sanitation Management 569
56. From Textual Variation to Functional Diversity in Beginner C Programming Tasks Generated by Ollama Cloud-Based Large Language Models 583
57. Impact of AI Coding Agents on Software Engineering with Junior Developers: An Experiment in Situ 597
58. HID Hardening vs. FIDO2: A Security-versus-Security Trade-off 607
59. AI in Decision-Making: The Case of University Metropolitan Tirana 615
60. Emerging Technology Adoption in TVET: A Theoretical and Pedagogical Synthesis 625
61. Randomized Backtracking: Another Unusable Sampling Method for the Traveling Tournament Problem? 634
62. Educational Research in Computer Science: Data Collection, Processing, and Analysis 642
63. Siamese Swin Transformers for Robust Polyp Classification 651

64. Architecture of a Fuzzy Logic Model for Supporting Investigations in Burglary Cases 663
65. Secure Biometric Data Management On A Permissioned Blockchain: A Hyperledger Besu-Based Architecture 673
66. Reinforcement Learning-Based Adaptive Hybrid PSO-GA Framework for Solving Dynamic Optimization Problems 683
67. Assessing the Organizational Adoption of Generative AI: Algorithmic Benchmarking and Empirical Insights into Corporate Transformation 689
68. A Tabletop Exercise Framework for Cross-Border Digital Health Incident Response 707
69. A Multimodal Analysis Approach for Automated Classification of Content Rating Symbols in Video Content 717
70. AI-Supported Emission Factor-Based Approach for Carbon Footprint Analysis and Intersection Prioritization in Sustainable Cities 725
71. Multi-Source Behavior-Based Anomaly Detection for Smart Water Management Networks Using SNMP and Syslog Telemetry 735
72. A Novel PL/SQL-Based Data Transformation and Optimization Framework for Enterprise Data Warehouses 744
73. Multi-Agent Conflict Simulation With World Constraints (MACSWC) 754
74. Trustworthiness Under Compression: Calibration and Explanation Faithfulness of Lightweight IoT Intrusion Detectors 762
75. A Dataset-Based Exploratory Study of Cybersecurity Threat Patterns from 2015 to 2024 771
76. Ability-Based Approaches for Ambient Assisted Living in Smart Homes 780
77. Various Topological Graph Indices In Sierpinski Triangle Networks 791
78. HeritageCare: A Predictive Maintenance And Monitoring System for Heritage Buildings 803
79. AurIA AI for University Administrative Systems 814
80. MuseDSL: A Domain-Specific Language and Federated Framework for Explainable Museum Object Linking 825
81. The Effects of Meta-Heuristic Optimization Algorithms on the Performance of Machine Learning Models for Alzheimer's Diagnosis 838
82. An NLP-Based Review and Taxonomy of Heart Murmur Detection Algorithms (PhysioNet 2022) 848
83. Enabling Clinical-Grade Left Ventricular Hypertrophy Detection in Wearable Devices via Deep Learning 858
84. Enhancing Breast Cancer Screening in Kenya Through Deep Learning-Driven Mammographic Analysis 863
85. Territorial Crime Pattern Classification in Peru Using Unsupervised Learning and K-Medoids, 2015–2024 877
86. Optimized Algorithm towards Addressing the Cold-Starts Issue in the Airline Recommender System 892
87. The Impact of Syntactic Dependency Information on Turkish Coreference Resolution 906
88. A Physics-Informed Hybrid Framework for Predictive Supervision and Multi-Objective Optimization of Ultrafiltration Membrane Fouling in Desalination 916
89. Leakage-Aware and Explanation-Stable XAI for Short-Term Smart Home Appliance Energy Forecasting 925
90. A Work-in-Progress Cross-Case Analysis of Three VR Exposure Prototypes 934
91. Semantic Interoperability of Heterogeneous Research Data: A FAIR-Aligned Ontology-Driven Framework for Interdisciplinary Research 941
92. A Comparative Analysis of Optimal Curtailment Strategies for Distributed Energy Resources 962
93. TRITON: A Modular Transformer for Per-Pixel Marine Pollutant Detection in Sentinel-2 Imagery 975
94. Research Program: Computer-Aided Molecular Target Identification for Tuftsin-Phosphorylcholine in Autoimmune Diseases 989

95. Design of a Compact Asymmetric Biconical Antenna for UAV Applications 998
96. Leveraging AI Agents for Collaborative and Independent Authorship 1008
97. Utilization of Honeypot Systems for Enhancing Cybersecurity in Smart Cities 1019
98. Systematic Analysis of Honeypot Deception Techniques: State-of-the-Art and Empirical Validation 1032
99. Influence of Acquisition Trajectory on Area Target Recognition and Crime Scene Re-Localization Using Consumer LiDAR 1046
100. The Dependency of NIS2 Critical Sectors on Operational Technology: Implications for Cybersecurity Resilience and Critical Infrastructure Protection 1057
101. Learnscrolling: Harnessing Attention Economy Mechanics for Incidental Learning via Synthetic Social Media 1070
102. From NIS to NIS2: Strengthening Cybersecurity for Critical Infrastructure Protection in the European Union 1080
103. Computational Models for Predicting Female Breast Cancer Survival and an Onco-Hospital Feature Engineering Workflow 1093
104. Precise Segmentation of Dental Implants with YOLO-Based Deep Learning Models 1103
105. PhysioReco: An Evidence-Guided Machine Learning System for Personalized Arthritis Exercise Recommendation 1113
106. Blockchain-Based Associate Professor Evaluation System: Transparent Academic Career Management with Ethereum Smart Contracts and IPFS 1131
107. A Region-Ablation Study of MRI Texture Features for Trigeminal Neuralgia Classification 1141
108. From Classical Retrieval to LLM-Native and Multimodal Search: A Synthesis of an Ongoing Convergence 1149
109. Bibliometric Analysis of Artificial Intelligence Assisted Medical Imaging Research: Global Trends and Turkey's Position (2021-2025) 1160
110. aBERT-Based Hybrid Models for Fake News Detection: A Systematic Review from 2021 to 2026 1172
111. Using Long Short-Term Memory (LSTM) Algorithm to Prevent the Foreign Currency Speculation in Libyan Black Markets 1183
112. Towards Multimodal Artificial Intelligence (OMEAI) Framework for Early Detection of Postnatal Depression 1196
113. Development of a Novel Low-Time-Complexity Optimization Technique for Edge-Based Wireless Body Area Networks 1211
114. Comparative Analysis of Evolutionary, Swarm, and Lamarckian Memetic Algorithms for Route Optimization in Industrial Printed Circuit Board (PCB) Drilling Processes 1221
115. Oil Spill Segmentation in SAR Imagery Using Swin Transformer with Compound Loss: A Multi-Sensor Evaluation 1227
116. FedSLM: A Federated Learning Approach Based on Fine-Tuning a Small Language Model for Text Classification 1240
117. Smart Route: A Road-Network-Aware Trip Evaluation and Ranking Microservice for Destination-Bound Ride-Hailing Drivers 1251
118. FSLM-SQL: A Novel Framework Based on Fine-Tuning Small Language Models for Natural Language to SQL 1260
119. Empirical Evaluation and Mathematical Limitations of PDE-Based Smoothing in Keystroke Dynamics Authentication 1271
120. A Seven-Category Error Taxonomy for Implicit Aspect Detection: Systematic Failure Analysis of BERT and GPT-2 in Aspect-Based Sentiment Analysis 1277
121. A Context-Aware Computer Vision-Based Model for ACL Injury Prevention in Female Youth Football: A PRISMA-based Review 1290
122. Synthetic-to-Real Camera Contamination Classification Using Modern Deep Neural Networks 1304

123. Evaluating TOE Determinants of Big Data Analytics Readiness through CRITIC–TOPSIS: Evidence from Albania 1315
124. Traffic Accident Severity Prediction with an Interpretable Crash Severity Risk Score 1325
125. Cloud-CTI-Engine: An Integrated AWS Security Posture and Threat Intelligence Platform 1335
126. Context-Aware Regular Expressions for Sensitive Data Detection 1344
127. Privacy-Preserving Secret Validation: An Empirical Comparison of Probabilistic Filters, k-Anonymity, and Private Set Intersection 1352
128. Evaluating the Impact of Data Augmentation on Cross-Dataset Generalization in Intrusion Detection Systems 1362
129. Selective Data Cleaning with Large Language Models for Improving E-Commerce Product Matching Training Data 1372
130. AI-Powered Smart Sensor for Direct Analysis of Aflatoxins in the Open Air Environment 1382
131. Neuro-Symbolic Correction of Kazakh Word Forms through Morphological Analysis and Selective Abstention 1392
132. Non-intrusive Human Activity Recognition using Imbalanced mm-Wave Doppler Radar Streams and Ensemble of Time-Series Modeling Strategies 1410
133. Full-Frame Sparse Temporal Multitask Learning for Student Affective State Recognition Using DAiSEE 1419
134. Visual Editing and Debugging of Feature Models 1428
135. Issue Analyzer: Similarity-Based Grouping and Team Routing for Mobile Crash Logs 1435
136. Design Science Research for Digital Platforms in Resource-Constrained Environments: The Lumikiza Case 1445
137. Brain-Computer Interface, Brain Science, and Artificial Intelligence 1462
138. Scalability Analysis of CNN-Based sEMG Hand Gesture Recognition Across Increasing Gesture Repertoires 1469
139. Integrating NotebookLM into Flipped Mathematics Instruction: Insights from Pre-Service Teachers through the TPACK Framework 1478
140. SpecGuard: Multi-Layer Quality Assessment of Natural-Language FPGA Requirements 1488
141. Counterfactual Explanations for Gallstone Prediction: A Comparative Evaluation of DiCE and PertCF 1499
142. Classification of Eight Different Retinal Conditions in Retinal OCT Images Using Convolutional Neural Networks 1508
143. AI-assisted Error Analysis to Design a Writing Course Syllabus 1516

CONFERENCE PROGRAM

17 September 2026 08:00 – 09:00 GMT+2	**KEYNOTE SPEECH** **Title:** Impact of Blockchain on Key Enabling Technologies for the Metaverse **Prof. Dr. Thippa Reddy Gadekallu** Professor at Zhejiang A&F University Chief Engineer at Zhongda GroupHaiyan County, Jiaxing City, Zhejiang Province, China Session Chair: Assoc. Prof. Dr. Adem Tekerek

Hall	**FUZULI**
Session	**1**
Topic	**Artificial Intelligence, Machine Learning & Applied AI**
Session Chair	**Assoc. Prof. Dr. Adem Tekerek (Gazi University)**
Date	**17 September 2026**
Time	**09:00 – 14:00 GMT+2**

	Time	#ID	Presentation Title	Authors
1	09:00 – 09:10	#28 Azerbaijan	Innovative yet Low-Resource Uses of AI in Teaching: Pedagogical Boundaries and Ethical Considerations for Teachers	***Mehriban A. Hasanova*** *(Ganja State University)*
2	09:10 - 09:20	#37 Kazakhstan	A methodological framework for preparing future computer science teachers for sustainable professional activity in the artificial intelligence context	***Aizhan Adambekova****, Aigul Aldabergenova, Aiym Yesseikyzy, Ulzhalgas Yesseikyzy, Gulmira Kasteyeva (Zhetysu University)*
3	09:20 – 09:35	#50 México	Machine Learning in Agile Project Management: A Bibliometric Analysis	***Yadira Jazmín Pérez Castillo*** *(Instituto Politécnico Nacional), Sandra Dinora Orantes Jiménez (Instituto Politécnico Nacional)*
4	09:35 – 09:45	#59 Kazakhstan	The Illusion of Cultural Competence: Evaluating ChatGPT-Generated Kazakh Proverbs	***Shynar Auyelbekova*** *(SDU University), Gulbanu Mukhiyayeva*
5	09:45 – 09:50	#69 India	From Judgment to Benchmark: An AI-Assisted Wedge Classification Framework for Commercial Loan Spread Governance	***Anandasubramanian Pranatharthy Codangudi*** *(VELS University)*
6	09:50 – 10:05	#85 North Macedonia	Artificial Intelligence in Online Gambling and Betting Systems: A Knowledge Representation–Based Use Case	***Trajance Angelovski****, Toni Malinovski, Sime Arsenovski, Miodrag Cekikj (University of Skopje)*

7	10:05 – 10:20	#86 México	Spatiotemporal pattern detection of precipitation using machine learning and remote sensing data *Ramón Fernando López Osorio, María Alejandra Quintero Morales, Yedid Guadalupe Zambrano-Medina* ***Lidia Yadira Pérez-Aguilar,** Evangelina Avila-Aceves, Sergio Alberto Monjardin-Armenta (Universidad Autónoma de Sinaloa)*
8	10:20 – 10:30	#90 Georgia	Boundary Contact Modeling of Tumor Growth with Neural Surrogate Acceleration for Embedded Biomedical Systems ***Manana Chumburidze**, Teimuraz Sakhelashvili, Giorgi Chachua, Nona Kukhianidze (Akaki Tsereteli State University)*
9	10:30 – 10:45	#97 Türkiye	A Comparative Analysis of Vision Transformer Models for Multimodal Emotion Recognition ***İlya Kuş**, Cemal Koçak, Ayşe Keleş (Gazi University)*
10	10:45 – 10:55	#102 Türkiye	Deep Learning Model and Zonal Clustering for RNA 3D Structure Prediction ***Mustafa Kemal Binli**, Adem Tekerek, Necaattin Barışçı (Gazi University)*
11	11:30 – 11:40	#104 United Kingdom	Explainable Multi-Type Loneliness Detection from Reddit Using Classical ML, Neural Networks, and Transformer Models with LIME and SHAP Analysis ***Alif Laila**, Arif Ahmed Dar, Hamidreza Khaleghzadeh (University of Portsmouth)*
12	11:40 - 11:55	#123 Azerbaijan, Russia	Artificial Intelligence in Business: Opportunities and Risks Sergey Bogatenkov, Irina Ketova, **Heyran Abbasova**, Nizami Yusubov, Mukhaddin Samadov
13	11:55 – 12:10	#138 Bosnia and Herzegovina	Explainable Machine Learning for Heart Disease Risk Prediction: Comparative Evaluation of ML Algorithms, Feature Importance Stability, and Clinical Interpretability ***Medina Ferhatović**, Belmin Haračić, Narcisa Hadžajlić (University of Zenica)*
14	12:10 – 12:30	#143 Kazakhstan	Evaluating the Effectiveness of the AIM-5 Model for AI Teaching Self-Efficacy and AI Integration Readiness ***Madina Bokan**, Zhomart Zhiyembayev, Mansur Tokanov, Damira Shalkarkyzy (Zhetysu University)*
15	12:30 – 12:45	#144 Kuwait	Generative AI through the Student Eye: perspectives, attitudes, and usage ***Mariam Alshammari**, Khuloud Alkhateeb, Fajer Alshammari, Fatma Alalawi (Gulf University for Science and Technology)*
16	12:45 – 13:00	#151 Albania, USA	AI in Decision-Making: The Case of University Metropolitan Tirana ***Suada Gjimi**, Arben Asllani (University Metropolitan Tirana)*
17	13:00 – 13:15	#162 Israel	Assessing the Organizational Adoption of Generative AI: Algorithmic Benchmarking and Empirical Insights into Corporate Transformation ***Idan Roth** (The Academic College of Tel Aviv-Yaffo)*
18	13:15 – 13:30	#177 México	AurIA: AI for University Administrative Systems **Sergio Yorick Álvarez Montalvo**, Damián Ulises Cedillo Rodríguez, Cindy Nallely Muñiz Pérez (*Universidad Autónoma de Tamaulipas)*
19	13:30 – 13:45	#183 Türkiye	An NLP-Based Review and Taxonomy of Heart Murmur Detection Algorithms (PhysioNet 2022) ***Onuralp Arslan** (Izmir Katip Celebi University), Yalçın İşler (Alanya Alaaddin Keykubat University)*
20	13:45 – 14:00	#185 Kenya	Enhancing Breast Cancer Screening in Kenya Through Deep Learning-Driven Mammographic Analysis *Edna C. Too, Saif Kinyori, David G Mwathi, Julius Gachoki, **Pauline W. Mwaka**, Mary Wanjiku, Lucy Gitonga (Chuka University)*

Hall	**FUZULI**
Session	**2**
Topic	**Artificial Intelligence, Machine Learning & Applied AI**
Session Chair	**Assoc. Prof. Dr. Adem Tekerek (Gazi University)**
Date	**18 September 2026**
Time	**08:30 – 11:00 GMT+2**

		#ID	Presentation Title	Authors
21	08:30 - 8:50	#192 Libya, Nigeria	Optimized Algorithm towards Addressing the Cold-Starts Issue in the Airline Recommender System	*Ayman Mahmud (Aljafara University), Zakaria Suliman Zubi (Sirte University), Asma Alhadi Billayl (Aljafara University),* ***Adud Musa Mabu*** *(Yobe State University)*
22	08:50 - 09:05	#193 Türkiye	The Impact of Syntactic Dependency Information on Turkish Coreference Resolution	***Ebru Çıracı****, Nesrin Aydın Atasoy (Karabük University)*
23	09:05 - 09:20	#205 USA	Leveraging AI Agents for Collaborative and Independent Authorship	*Lamont Grissom (University of Arkansas at Little Rock), Daniel Berleant (University of Arkansas at Little Rock),* ***Venkata Batthula*** *(Southwestern College Winfield), Richard S. Segall (Arkansas State University Jonesboro)*
24	09:20 – 09:35	#214 Nigeria	PhysioReco: An Evidence-Guided Machine Learning System for Personalized Arthritis Exercise Recommendation	*Theresa Omodunbi,* ***Adesola Kareem****, Adekemi Amoo, Temitope Sholanke, Grace Alilu, Adaobi Okonji*
25	09:35 – 09:45	#219 Senegal	From Classical Retrieval to LLM-Native and Multimodal Search: A Synthesis of an Ongoing Convergence	*Fatou Kamara,* ***Mouhamadou Lamine Diakhame*** *(Gaston Berger University)*
26	09:45 - 10:05	#225 Libya	Using Long Short-Term Memory (LSTM) Algorithm to Prevent the Foreign Currency Speculation in Libyan Black Markets	***Zakaria Suliman Zubi*** *(Sirte University), Zakariya A Alsharef (Libyan Academy Misurata), Abdulbsit M Alajnaf (Higher Institute of Science and Technology Misurata)*
27	10:05 – 10:20	#227 United Kingdom	Towards Multimodal Artificial Intelligence (OMEAI) Framework for Early Detection of Postnatal Depression	***Gleslie Sybil Ann A. Maraon****, Xiaohong W. Gao (Middlesex University)*
28	10:20 – 10:35	#235 Vietnam	FedSLM: A Federated Learning Approach Based on Fine-Tuning a Small Language Model for Text Classification	**Nhat-Thien Nguyen**, Quang-Hung Le (*Quy Nhon University*)
29	10:35 – 10:50	#240 Vietnam	FSLM-SQL: A Novel Framework Based on Fine-Tuning Small Language Models for Natural Language to SQL	***Khac-Lap Phan****, Quang-Hung Le (Quy Nhon University)*

Hall	**FUZULI**
Session	**3**
Topic	**Artificial Intelligence, Machine Learning & Applied AI**
Session Chair	**Assoc. Prof. Dr. Diana Bratić (University of Zagreb)**
Date	**18 September 2026**
Time	**11:00 – 14:10 GMT+2**

		#ID	Presentation Title	Authors
30	11:00 – 11:40	#246 Malaysia	A Seven-Category Error Taxonomy for Implicit Aspect Detection: Systematic Failure Analysis of BERT and GPT-2 in Aspect-Based Sentiment Analysis	***Imelda Pangaribuan**, Arifah Che Alhadi, Mohamad Nor Hassan, Masita Jalil (Universiti Malaysia Terengganu)*
31	11:40 – 11:50	#259 Türkiye	Selective Data Cleaning with Large Language Models for Improving E-Commerce Product Matching Training Data	***Oğuz Akif Tüfekcioğlu** (Cimri Inf. Tech. Syst. Inc.), Özgür Aslan (Cimri Inf. Tech. Syst. Inc.), Tuna Çakar (MEF University)*
32	11:50 – 12:05	#261 Kazakhstan	Neuro-Symbolic Correction of Kazakh Word Forms through Morphological Analysis and Selective Abstention	*Laura Baitenova (Turan University), Saule Tussupova (Turan University), Gulnar Mukhamejanova (Narxoz University), Akhmed Baikhojayev (Narxoz University), **Gauhar Munaitbas** (Home Credit Bank JSC), Daryn Nurtaza (Turan University)*
33	12:10 – 12:20	#280 Serbia	Integrating NotebookLM into Flipped Mathematics Instruction: Insights from Pre-Service Teachers through the TPACK Framework	***Aleksandar Milenković**, Nemanja Vučićević, Marina Svičević (University of Kragujevac)*
34	12:20 – 12:35	#41 Türkiye	AI-Driven Dynamic Trust Scoring over Self-Sovereign Identity System	***Deniz Salucu** (TURKCELL), Tunga Sayıcı (TURKCELL)*
35	12:35 - 12:50	#45 India	Evaluating Quantum Computing Use Cases in Banking: A Sequential Model-Risk Admissibility Framework	**Anandasubramanian Pranatharthy Codangudi** (VELS University)
36	12:50 – 13:05	#49 Türkiye	A New Topological Index-Based Solution Approach for the Traveling Purchaser Problem	*Aysun Yurttaş Güneş, **İlker Küçükoglu**, Hacer Özden Ayna, İsmail Naci Cangül (Bursa Uludağ University)*
37	13:05 - 13:20	#54 Venezuela	Efficient Quantum Coherence/Decoherence Emulation Using Parallel-Recursive Computation on FPGAs and Algorithmic Skeletons	***Carlos Alfonso Acosta-León** (Central University of Venezuela)*
38	13:20 - 13:35	#57 Netherlands	Contemplating the Effects of NP-Complete Reductions on Problem Instance Hardness	***Daan van den Berg** (VU University Amsterdam), Thomas Meeus (University of Amsterdam), Joeri Sleegers (Mice & Man Software and A.I. Development), Pieter Adriaans (Institute for Logic, Language, and Computation)*
39	13:35 - 13:50	#60 Türkiye, Kazakhstan	A QUBO-Based Reformulation and Quantum Solution Framework for Nonlinear Constrained Optimization Problems	***Cumali Yaşar** (Çanakkale Onsekiz Mart University), Yerlan Izbassarov (Khalel Dosmukhamedov Atyrau University)*
40	13:50 – 14:10	#148 Germany	Impact of AI Coding Agents on Software Engineering with Junior Developers: An Experiment in Situ	***Sebastian Steger**, Kerstin Fink, Lukas Nock, Sven-Gunnar Klaus (Technische Hochschule Mannheim Mannheim)*

Hall	**GAZALİ**
Session	**1**
Topic	**Cybersecurity, Privacy, Digital Trust & Theoretical Computer Science**
Session Chair	**Prof. Dr. Ömer Faruk Bay (Gazi University)**
Date	**17 September 2026**
Time	**09:00 – 11:00 GMT+2**

	Time	#ID	Presentation Title	Authors
41	09:00 – 09:15	#61 Oman	Long Distance Environmental and Security System	***Mohammad Mohatram**, M. Khalfan Mohammed Al Hudaifi, Muhammad Irfan Memon, Javeed Hussain Shaik (Global College of Engineering and Technology)*
42	09:15 - 09:30	#66 North Macedonia	Comparative Evaluation of Heuristic Search Algorithms for Reward- Constrained Museum Visitor Routing on a Real-World Graph	***Petar Mileski**, Dijana Capeska Bogatinoska (University of Information Science and Technology "St. Paul the Apostle")*
43	09:30 - 09:45	#67 Croatia	Cybersecurity awareness and threat perception constructs	***Katarina Kostelić**, Dario Kukuljan, Mieta Bobanović Dasko (Juraj Dobrila University of Pula)*
44	09:45 - 10:00	#77 Bosnia and Herzegovina	From Regulatory Compliance to Digital Trust: Integrating NIS2 and ISO/IEC 27701 In Transitional Cybersecurity Governance - The Case of Bosnia and Herzegovina	***Edin Alić** (HIFA-OIL d.o.o. Tešanj), Muhamed Ćosić (Centar za održivost i digitalizaciju Centar EkoDigitus), Venan Hadžiselimović (HIFA-OIL d.o.o. Tešanj)*
45	10:00 – 10:10	#87 Georgia	Beyond Aggregate Metrics: Population-Weighted CVaR Auditing for Hidden Performance Failures Along Continuous Attributes	***Irakli Koberidze** (Tbilisi State University), Magda Tsintsadze (Tbilisi State University)*
46	10:10 - 10:15	#89 Türkiye, Ukraine	A Dual-Explainer GRU–TCN for Cyber-Physical Attack Detection in IoT-Based Energy Infrastructure	*Ahmet Nusret Özalp (Kastamonu University), **Viktoriia Hreniukh** (Ukrainian National Forestry University)*
47	10:15 – 10:25	#98 United Kingdom	Intelligent Cyber Threat Detection at Scale: A Comparative Study of Machine Learning and Deep Learning Approaches for Network Intrusion Detection on CICIDS2017	***Md Anas Biswas** (University of Portsmouth), Syed Anisur Rahman (University of Portsmouth), Hamidreza Khaleghzadeh (University of Portsmouth)*
48	10:25 – 10:40	#136 Finland	Role-Based Phishing to Ransomware Readiness for Clinical Staff	*Leon Carpelan, Anderson Siqueira Teixeira, Wimpie Roothe, Robert Simonsuuri, **Jyri Rajamäki** (Laurea University of Applied Sciences)*
49	10:40 – 10:55	#163 Finland	A Tabletop Exercise Framework for Cross-Border Digital Health Incident Response	*Niko Haanpää, Teemu Koski, Tommi Elg, Ville Kivistö, **Jyri Rajamäki** (Laurea University of Applied Sciences)*

Hall	**GAZALİ**
Session	**2**
Topic	**Cybersecurity, Privacy, Digital Trust & Theoretical Computer Science**
Session Chair	**Prof. Dr. Erol Kurt (Gazi University)**
Date	**17 September 2026**
Time	**12:00 – 15:20 GMT+2**

	Time	#ID	Presentation Title	Authors
50	12:00 – 12:15	#141 Hungary, Portugal	From Neo4j to Redis: Storing Graph Data and Comparing Graph Algorithm Performance	***Anikó Vágner*** *(University of Debrecen), Fanni Belics (University of Debrecen), Cristina Maria Costa Rodrigues (Polytechnic Institute of Santarém)*
51	12:15 – 12:30	#145 Azerbaijan	Analysis of Existing Approaches in The Field of Cryptographic Protocols in UAVs	***Ilaha Gahramanova*** *(Azerbaijan Technical University)*
52	12:30 – 12:45	#150 Azerbaijan	HID Hardening vs. FIDO2: A Security-versus-Security Trade-off	***Emin Huseynov*** *(Azerbaijan Technical University)*
53	12:45 – 13:00	#154 Netherlands	Randomized Backtracking: Another Unusable Sampling Method for the Traveling Tournament Problem	*Timothy Böhm (VU University), Daan van den Berg (VU University),* ***Bas Loyen*** *(VU University)*
54	13:00 – 13:15	#209 Bosnia and Herzegovina	The Dependency of NIS2 Critical Sectors on Operational Technology: Implications for Cybersecurity Resilience and Critical Infrastructure Protection	***Anes Mirojević*** *(University of Buckingham), Jasmin Azemović (University of Buckingham)*
55	13:15 – 13:25	#171 United Kingdom	Trustworthiness Under Compression: Calibration and Explanation Faithfulness of Lightweight IoT Intrusion Detectors	*Md Anas Biswas (University of Portsmouth,* ***Alif Laila*** *(University of Portsmouth)*
56	13:25 – 13:40	#172 North Macedonia	A Dataset-Based Exploratory Study of Cybersecurity Threat Patterns from 2015 to 2024	***Hristina Oracheva****, Dijana Capeska Bogatinoska, Daniela Mechkaroska (University of Information Science & Technology St. Paul the Apostle)*
57	13:40 – 14:50	#211 Bosnia and Herzegovina	From NIS to NIS2: Strengthening Cybersecurity for Critical Infrastructure Protection in the European Union	***Anes Mirojević*** *(University of Buckingham), Jasmin Azemović (University of Buckingham)*
58	14:50 – 15:10	#206 Czech Republic	Utilization of Honeypot Systems for Enhancing Cybersecurity in Smart Cities	*Vladimir Sobeslav (University of Hradec Kralove),* ***Daniel Kucera*** *(University of Hradec Kralove), Jan Pulpan (University of Hradec Kralove)*
59	15:10 – 15:20	#207 Czech Republic	Systematic Analysis of Honeypot Deception Techniques: State-of-the-Art and Empirical Validation	*Vladimir Sobeslav (University of Hradec Kralove),* ***Daniel Kucera*** *(University of Hradec Kralove), Jan Pulpan (University of Hradec Kralove)*

Hall	**GAZALİ**
Session	**3**
Topic	**Cybersecurity, Privacy, Digital Trust & Theoretical Computer Science**
Session Chair	**Prof. Dr. Ayhan Erdem (Gazi University)**
Date	**18 September 2026**
Time	**08:30 – 10:30 GMT+2**

	Time	#ID	Presentation Title	Authors
60	08:30 – 08:45	#166 Türkiye	Multi-Source Behavior-Based Anomaly Detection for Smart Water Management Networks Using SNMP and Syslog Telemetry	***Dilara Çetin**, Ayşegül Kara, Deniz Balta, Musa Balta (Sakarya University)*
61	08:45 – 09:00	#175 Türkiye	Various Topological Graph Indices in Sierpinski Triangle Networks	*İpek Akdoğan, Aysun Yurttaş Güneş, Hacer Özden Ayna, Aykut Şen, İlker Küçükoğlu, **İsmail Naci Cangül** (Bursa Uludag University)*
62	09:00 - 09:15	#217 Türkiye	Blockchain-Based Evaluation System: Transparent Academic Career Management with Ethereum Smart Contracts and IPFS	***Kemal Akkanat**, Muhittin Tanoba (Trakya University)*
63	09:15 - 09:30	#245 Türkiye	Empirical Evaluation and Mathematical Limitations of PDE-Based Smoothing in Keystroke Dynamics Authentication	***Levent Yalçın** (Acıbadem Mehmet Ali Aydınlar University), Ramazan Subaşı (Yıldız Teknik University)*
64	09:30 - 09:40	#251 Kosova	Cloud-CTI-Engine: An Integrated AWS Security Posture and Threat Intelligence Platform	***Gerta Hodolli**, Rrezearta Thaqi, Blerim Rexha (University of Prishtina)*
65	09:40 - 09:50	#254 Türkiye	Context-Aware Regular Expressions for Sensitive Data Detection	***Enes Can Güven** (Trendyol Group), Burak Tahtacı (Trendyol Group), Tuna Çakar (MEF University)*
66	09:50 – 10:00	#255 Türkiye	Privacy-Preserving Secret Validation: An Empirical Comparison of Probabilistic Filters, k-Anonymity, and Private Set Intersection	***Burak Tahtacı** (Trendyol Group), Enes Can Güven (Trendyol Group), Tuna Çakar (MEF University)*
67	10:00 - 10:15	#257 Algeria	Evaluating the Impact of Data Augmentation on Cross-Dataset Generalization in Intrusion Detection Systems	***Mounira Zekiouk**, Madjed Bencheikh Lehocine, Ahlam Bouanane (University of Mila)*
68	10:15 – 10:30	#160 Türkiye	Secure Biometric Data Management On A Permissioned Blockchain: A Hyperledger Besu-Based Architecture	***Durmuş Gülbahar**, Ahmet Gürsel, Ali Hakan Işık (Burdur Mehmet Akif Ersoy University)*

Hall	**NICOLA TESLA**
Session	**1**
Topic	**Computer Vision, Data Science, IoT & Edge Computing**
Session Chair	**Assoc. Prof. Dr. Eda Akman Aydın (Gazi University)**
Date	**17 September 2026**
Time	**09:00 – 11:30 GMT+2**

	Time	#ID	Presentation Title	Author
69	09:00 – 09:10	#56 Algeria	Mining Sequential Rules from Large-Scale Sequence Databases Using an FPTree-Based Method	***Imane Seddiki*** *(Mohamed El Bachir El Ibrahimi University), Abdelbasset Barkat (University of M'sila)*
70	09:10 – 09:20	#70 Türkiye	Evolution in Remote Sensing Change Detection: Methods, Challenges, and Next AI Perspectives	***Zeynep Açıkgöz*** *(Düzce University), İbrahim Alper Doğru (Gazi University), Kenan Toklu (Namık Kemal University)*
71	09:20 – 09:30	#76 Kazakhstan	Enhancing Small Object Detection in Mammography Using a Patch-Based YOLO Framework	***Aray Aryn*** *(Kazakh National Research Technical University), Orken Mamyrbayev (Institute of Information and Computing Technologies), Turdybek Kurmetkan (Al-Farabi Kazakh National University)*
72	09:30 - 09:45	#81 Canada	Adaptive Lossless Compression for Event-Based Cameras	*Prottasha D'Cruze,* ***Howard Cheng*** *(University of Lethbridge)*
73	09:45 – 10:00	#91 Georgia	Decision Tree-Based Data Completion in Structured Lexicographic Resources	***Archil Elizbarashvili*** *(Tbilisi State University), Magda Tsintsadze (Tbilisi State University)*
74	10:00 – 10:15	#108 Türkiye	Water Quality Monitoring and Pollution Detection Using Edge-IoT-Based Hybrid Artificial Intelligence Models: A Systematic Literature Review	***Hüseyin Özgür Kazan****, İsmail Atacak, İbrahim Alper Doğru (Gazi University)*
75	10:15 – 10:25	#109 Türkiye	A Practical Evaluation of CNN Architectures for Histopathological	***Nazlı Çıtak*** *(Gazi University), Ömer Faruk Bay (Gazi University)*
76	10:25 – 10:45	#115 Kazakhstan	Edge-Based Digital Twin for Real-Time Energy Monitoring in Food Manufacturing: A Case Study of a Bakery Enterprise	*Bauyrzhan Amirkhanov, Murat Kunelbayev, Miras Tokhtassyn,* ***Alina Raeva****, Gulshat Amirkhanova (Al-Farabi Kazakh National University)*
77	10:45 – 11:00	#116 Kazakhstan	Machine Learning Models for Monitoring and Analysis of Mobile Object Movements at a Food Industry Enterprise	***Bauyrzhan Amirkhanov****, Yenlik Faruzkyzy, Alikhan Amirkhanov, Gulshat Amirkhanova, Dina Baizhanova (Al-Farabi Kazakh National University)*
78	11:00 – 11:10	#121 Türkiye	Explainable Mammography Classification via Attention-Guided SE-EfficientNet-B3 and Grad-CAM Diagnostic Interpretation	***Abdine Ben Ali****, Sinan Toklu (Gazi University)*
79	11:10 – 11:30	#132 Libya	Harnessing Data Science and Machine Learning for Financial Fraud Detection Using the Central Bank of Libya Dataset	***Zakaria Suliman Zubi*** *(Sirte University), Abdulbasit M Alajnaf (Higher Institute of Science and Technology Misurata, Zakariya A Alsharef (Libyan Academy, Misurata)*

Hall	**NICOLA TESLA**
Session	**2**
Topic	**Computer Vision, Data Science, IoT & Edge Computing**
Session Chair	**Prof. Dr. Murat Dörterler (Gazi University)**
Date	**17 September 2026**
Time	**12:00 – 14:30 GMT+2**

	Time	#ID	Presentation Title	Author
80	12:00 – 12:15	#135 Cuba, Spain	Wavelet-Based Machine Learning for Real-Time Burst Detection on Embedded Edge Devices ***Jaime Ernesto Chiang Crúz*** *(Center for Hydraulic Research, University of Havana "José Antonio Echeverría"), Jorge Ramírez-Beltrán Center for Hydraulic Research, University of Havana "José Antonio Echeverría"), Sergio Enrique Rodríguez Pérez Center for Hydraulic Research, University of Havana "José Antonio Echeverría"), Alex Macia-Fiteni (University of Alicante), Iren Lorenzo-Fonseca (University of Alicante)*	
81	12:15 – 12:25	#140 México	Preliminary Development of an Aircraft Structural Damage Classification System Using Deep Learning ***Reynaldo Ángel Serrano Cruz****, Gloria Lilia Osorio Gordillo, Rodolfo Amalia Vargas Méndez (Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET))*	
82	12:25 – 12:35	#146 Nigeria	SHRUMAC: An IoT-Smart Hygiene, Refuse Monitoring and Alert Controller System for Real-Time Institutional Campus Sanitation Management ***Shobowale, Rukayat Temitope*** *(Federal University Otuoke), Maclean, Friday Maclean (Federal University Otuoke)*	
83	12:35 – 12:50	#156 Türkiye	Siamese Swin Transformers for Robust Polyp Classification ***Mesut Şeker*** *(Dicle University)*	
84	12:50 – 13:05	#164 Türkiye	A Multimodal Analysis Approach for Automated Classification of Content Rating Symbols in Video Content ***Barış Şenel****, Murat Dörterler, Adem Tekerek (Gazi University)*	
85	13:05 - 13:20	#174 Sweden	Ability-Based Approaches for Ambient Assisted Living in Smart Homes ***Marijana Teljega*** *(Kristianstad University), Daniel Einarson (Kristianstad University)*	
86	13:20 – 13:30	#176 Malaysia	HeritageCare: A Predictive Maintenance and Monitoring System for Heritage Buildings ***Chan Rou Mei*** *(TAR UMT), Cho Sook Fong (TAR UMT), Loke Sue Voon (TAR UMT)*	
87	13:30 – 13:45	#184 Türkiye	Enabling Clinical-Grade Left Ventricular Hypertrophy Detection in Wearable Devices via Deep Learning ***Arda Sarpay*** *(Izmir Katip Celebi University), Yalçın İşler (Alanya Alaaddin Keykubat University)*	
88	13:45 - 13:55	#186 Peru	Territorial Crime Pattern Classification in Peru Using Unsupervised Learning and K-Medoids, 2015–2024 ***Renzo Alberto Taco Coayla****, Martha Judith Paredes Vignola, Enrique Félix Lanchipa Valencia, Elard Ricardo Rodríguez Marca, Hugo Martín Alcántara Martínez, Josué Abraham Emanuel Chambilla Zúñiga* *(Universidad Privada de Tacna, Tacna, Perú)*	
89	13:55 - 14:10	#194 Morocco	A Physics-Informed Hybrid Framework for Predictive Supervision and Multi-Objective Optimization of Ultrafiltration Membrane Fouling in Desalination ***Zineb Idabella*** *(Cadi Ayyad University), Laila Amir (Cadi Ayyad University)*	

90	14:10 – 14:25	#195 New Zealand	Leakage-Aware and Explanation-Stable XAI for Short-Term Smart Home Appliance Energy Forecasting ***Ahmad Wedyan*** *(Manukau Institute of Technology), David McCurdy (Manukau Institute of Technology)*
91	14:25 – 14:40	#249 Albania	Evaluating TOE Determinants of Big Data Analytics Readiness through CRITIC–TOPSIS: Evidence from Albania ***Megi Marku*** *(Agricultural University of Tirana)*

Hall	**NICOLA TESLA**
Session	**3**
Topic	**Computer Vision, Data Science, IoT & Edge Computing**
Session Chair	**Assoc. Prof. Dr. Saadin OYUCU (Gazi University)**
Date	**18 September 2026**
Time	**08:30 – 12:30 GMT+2**

	Time	#ID	Presentation Title	Authors
92	08:30 - 08:50	#198 Uganda	Semantic Interoperability of Heterogeneous Research Data: A FAIR-Aligned Ontology-Driven Framework for Interdisciplinary Research	***Kevin Aber*** *(Gulu University), Geoffrey Olok Tabo (Gulu University), Benedict Oyo (Gulu University), Lone Dirckinck Holmfeld (Aalborg University), Christine Kalumera Akello (Gulu University), Daniel Ogenrwot (Gulu University)*
93	08:50 – 09:05	#202 Mexico	TRITON: A Modular Transformer for Per-Pixel Marine Pollutant Detection in Sentinel-2 Imagery	*José Luis de León Carmona (Center for Research in Mathematics (CIMAT),* ***Rogelio Hasimoto-Beltrán*** *(Center for Research in Mathematics (CIMAT))*
94	09:05 – 09:20	#212 Brazil, Portugal	Computational Models for Predicting Female Breast Cancer Survival and an Onco-Hospital Feature Engineering Workflow	***Daniela Schimitz de Carvalho*** *(University of Porto), Maximiliano Ribeiro Guerra (Federal University of Juiz de Fora), Marco Paulo Lages Parente (University of Porto), Priscila Vanessa Zabala Capriles Goliatt (Federal University of Juiz de Fora)*
95	09:20 – 09:35	#213 Türkiye	Precise Segmentation of Dental Implants with YOLO-Based Deep Learning Models	*Hüseyin POLAT (Gazi University), Emin UYSAL (Pi Health Technologies),* ***Hakan ÜNAL*** *(Gazi University), Elif Meltem ASLAN (University of Health Sciences)*
96	09:35 - 09:50	#218 Türkiye	A Region-Ablation Study of MRI Texture Features for Trigeminal Neuralgia Classification	***Önder Polat*** *(Gaziantep Islam Science and Technology University), Zeynep Ziroğlu (Ankara Training and Research Hospital), Habip Eser Akkaya (Ankara Training and Research Hospital), Hafize Nalan Güneş (Ankara Training and Research Hospital)*
97	09:50 – 10:05	#220 Türkiye	Bibliometric Analysis of Artificial Intelligence-Assisted Medical Imaging Research: Global Trends and Turkey's Position (2021-2025)	***Kemal Akkanat*** *(Trakya University), Muhittin Tanoba (Trakya University)*
98	10:30 – 10:45	#229 SriLanka, Malaysia	Development of a Novel Low-Time-Complexity Optimization Technique for Edge-Based Wireless Body Area Networks	***Sachinthani Alahakoon*** *(Gampaha Wickramarachchi University and Indigenous Medicine), Mustafa Muwafak Alobaedy (City University)*
99	10:45 -11:00	#234 Mexico	Oil Spill Segmentation in SAR Imagery Using Swin Transformer with Compound Loss: A Multi-Sensor Evaluation	*Jose Daniel Foronda Ortiz (Center for Research in Mathematics),* ***Rogelio Hasimoto Beltrán*** *(Center for Research in Mathematics)*
100	11:00 – 11:10	#248 Hungary	Synthetic-to-Real Camera Contamination Classification Using Modern Deep Neural Networks	***János Hollósi*** *(Széchenyi István University)*
101	11:10 – 11:25	#250 Türkiye	Traffic Accident Severity Prediction with an Interpretable Crash Severity Risk Score	***Sevcel Gözde BILGIN*** *(Gazi University), Cemal KOCAK (Gazi University)*

102	11:25 – 11:35	#260 Uganda	AI-Powered Smart Sensor for Direct Analysis of Aflatoxins in the Open Air Environment ***Fred Paul Mark Jjunju*** *(Universal Technology and Management University), Denis Bbosa (Mbarara University of Science and Technology)*
103	11:35 – 11:45	#262 United Kingdom	Non-intrusive Human Activity Recognition using Imbalanced mm-Wave Doppler Radar Streams and Ensemble of Time-Series Modeling Strategies ***Khadija Adam Nasir*** *(Swansea University), Sara Sharifzadeh (Swansea University), Alaa Humaidan (Swansea University), Jeny Roy (Swansea University)*
104	11:45 -11:55	#263 Türkiye	Full-Frame Sparse Temporal Multitask Learning for Student Affective State Recognition Using DAiSEE ***İbrahim Dolaş*** *(Gazi University), Hüseyin Polat (Gazi University)*
105	11:55 – 12:10	#247 South Africa	A Context-Aware Computer Vision-Based Model for ACL Injury Prevention in Female Youth Football: A PRISMA-based Review ***Zodzwainashe Kudyachete*** *(Nelson Mandela University), Chinonye Leuna Obioha (Nelson Mandela University), Darelle Van Greunen (Nelson Mandela University)*

Hall	**DA VINCI**
Session	**1**
Topic	**Software Engineering, HCI, Robotics & Applied Computing**
Session Chair	**Assoc. Prof. Dr. Sinan Toklu (Gazi University)**
Date	**17 September 2026**
Time	**09:00 – 11:00 GMT+2**

	Time	#ID	Presentation Title	Author
106	09:00 – 09:15	#38 Croatia	Gender Pay Gaps in Software Engineering: A Literature Review of Attrition, Measurement, and Structural Inequality ***Marija Macek*** *(University of Zagreb)*	
107	09:15 – 09:30	#42 Azerbaijan	Digital twin-based residual analysis for leak detection and localization in high-pressure gas pipelines ***Murad Sharifov*** *(Baku Higher Oil School), Naila Allahverdiyeva (Baku Higher Oil School)*	
108	09:30 – 09:45	#51 Oman	Real-Time Helmet Detection System with Monitoring and Alert for Restricted Areas ***Mohammad Mohatram****, Asma Abdallah Al Reesi, Javeed Hussain Shaik, Muhammad Irfan Memon (Global College of Engineering and Technology)*	
109	09:45 – 09:50	#58 Sri Lanka	Explainable Commit-Level Vulnerability Intelligence in DevSecOps Pipelines: A Systematic Review and Experimental Analysis ***Dasun Sandeepa Wilathgamuwa*** *(University of Sri Jayawardanapura), Malshani Heetiarachchi (University of Moratuwa)*	
110	09:50 – 10:00	#68 Türkiye	MPC-Based Trajectory Tracking and Dynamic Obstacle Avoidance for an Autonomous Vehicle in ROS2 ***Ibrahim Toy*** *(MPG Inc.), Akif Durdu (Konya Technical University)*	
111	10:00: 10:15	#73 Algeria	Multi-Objective Hybrid Swarm Optimization for Energy-Delay Trade-off in SDN-enabled MEC ***Cherhabil Fatima Z.*** *(Mustafa Benboulaid University (Batna2)), Bendib Sonia-Sabrina (Mustafa Benboulaid University (Batna2)), Sedrati Maamar (Mustafa Benboulaid University (Batna2))*	
112	10:15 - 10:30	#80 Ukraine	Reward System for Student Social Activity: Lntu Case Study ***Iurii Lukianchuk****, Lev Boyko, Yurii Tulashvili, Olga Nedilko (Lutsk National Technical University)*	
113	10:30 - 10:40	#107 Greece	Continuously Adaptive Digital Twins for Sim-to-Real Learning in Agro-Food Systems ***Georgios Protogeros****, Aikaterini Protogerou, Manos Roumeliotis (University of Macedonia)*	
114	10:40 - 10:50	#126 Türkiye	An Analysis of the Ministry of Health of the Republic of Türkiye from a Web Accessibility (Barrier-Free Design) Perspective ***Halime Suvay EKER*** *(İstanbul Gedik University)*	
115	10:50 – 11:00	#142 Spain, Cuba	Open cloud voice service architecture ***Alex Macia Fiteni*** *(University of Alicante), Jaime E. Chiang Cruz (Technological University of La Habana "José Antonio Echeverría"), Iren Lorenzo Fonseca (University of Alicante), Francisco Maciá Pérez (University of Alicante)*	
116	11:00 – 11:10	#147 Slovakia	From Textual Variation to Functional Diversity in Beginner C Programming Tasks Generated by Ollama Cloud-Based Large Language Models *Ladislav Végh (J. Selye University), Ondrej Takáč (J. Selye University),* ***Dávid Demeter*** *(J. Selye University),* *Krisztina Czakóová (J. Selye University)*	
117	11:10 – 11:20	#153 Zambia	Understanding Emerging Technology Adoption in Technical and Vocational Education and Training through Theoretical and Pedagogical Perspectives ***Twaambo Chiinza*** *(Technical and Vocational Teachers' College), Alice P. Shemi (Copperbelt University)*	
118	11:20 – 11:40	#155 Hungary	Educational Research in Computer Science: Data Collection, Processing, and Analysis ***Carolin Hannusch*** *(University of Debrecen), Mária Csernoch (University of Debrecen)*	

Hall	**DA VINCI**
Session	**2**
Topic	**Software Engineering, HCI, Robotics & Applied Computing**
Session Chair	**Assist. Prof. Dr. Halime Suvay EKER (İstanbul Gedik University)**
Date	**17 September 2026**
Time	**12:00 – 14:20 GMT+2**

	Time	#ID	Presentation Title	Author
119	12:00 - 12:10	#157 Bulgaria	Architecture of a Fuzzy Logic Model for Supporting Investigations in Burglary Cases ***Radmila Radostinova*** *Roydeva (Burgas Free University), Penka Georgieva (Burgas Free University)*	
120	12:10 – 12:25	#161 Uzbekistan	Reinforcement Learning-Based Adaptive Hybrid PSO-GA Framework for Solving Dynamic Optimization Problems ***Mageshkumar Chinnasamy*** *(Webster University), Nitin S. Goje (Webster University)*	
121	12:25 – 12:35	#165 Türkiye	AI-Supported Emission Factor-Based Approach for Carbon Footprint Analysis and Intersection Prioritization in Sustainable Cities ***Elif Yildiz*** *(İstanbul University-Cerrahpaşa), Deniz Balta (Sakarya University), Musa Balta (Sakarya University)*	
122	12:35 - 12:50	#167 USA	A Novel PL/SQL-Based Data Transformation and Optimization Framework for Enterprise Data Warehouses **Tirumalavenkata Naga Lakshmi Jammula**	
123	12:50 – 13:05	#168 North Macedonia	Multi-Agent Conflict Simulation With World Constraints (MACSWC) *Borjan Ladinski (University American College Skopje),* ***Dusko Sofronievski*** *(University American College Skopje),* *Marija Stankova Medarovska (University American College Skopje)*	
124	13:15 – 13:30	#210 Denmark	Learnscrolling: Harnessing Attention Economy Mechanics for Incidental Learning via Synthetic Social Media ***Edward Abel*** *(University of Southern Denmark)*	
125	13:30 - 13:45	#182 Türkiye	The Effects of Meta-Heuristic Optimization Algorithms on the Performance of Machine Learning Models for Alzheimer's Diagnosis ***Ömer Faruk Çakır****, Aysun Çoşkun, Adem Tekerek (Gazi University)*	
126	13:30 – 13:40	#197 Israel	A Work-in-Progress Cross-Case Analysis of Three VR Exposure Prototypes ***Marina Knyazhansky*** *(Shamoon College of Engineering)*	
127	13:40 - 13:55	#201 Türkiye	A Comparative Analysis of Optimal Curtailment Strategies for Distributed Energy Resources ***Mustafa Aker*** *(Siemens A.S.)*	
128	13:55 - 14:10	#204 Libya	Design of a Compact Asymmetric Biconical Antenna for UAV Applications *Ali Larbah, Salem Alkhair, Jalal Srar,* ***Hassan Aldeeb****, Mohammed Alzwawi (Misurata University)*	
129	14:10 – 14:20	#178 Türkiye	MuseDSL: A Domain-Specific Language and Federated Framework for Explainable Museum Object Linking ***Cumali Yaşar****, Zafer Karadayı, Ayten Çalık, Emin Ulugergerli (Çanakkale Onsekiz Mart University)*	

Hall	**DA VINCI**
Session	**3**
Topic	**Software Engineering, HCI, Robotics & Applied Computing**
Session Chair	**Asist Prof. Dr. Recep Başarıcı (Kahramanmaraş Sütçü İmam University)**
Date	**18 September 2026**
Time	**08:30 – 12:00 GMT+2**

		#ID	Presentation Title	Author
130	08:30 - 08:40	#224 Türkiye	BERT-Based Hybrid Models for Fake News Detection: A Systematic Review from 2021 to 2026 ***Dilara Çelik*** *(Gazi University), Necaattin Barişçi (Gazi University), Yusuf Sönmez (Gazi University)*	
131	08:40 - 08:55	#230 Türkiye	Comparative Analysis of Evolutionary, Swarm, and Lamarckian Memetic Algorithms for Route Optimization in Industrial Printed Circuit Board (PCB) Drilling Processes ***Ali Güner Bekir*** *(Gazi University), Aysun Coşkun (Gazi University)*	
132	08:55 - 09:10	#237 Türkiye	Smart Route: A Road-Network-Aware Trip Evaluation and Ranking Microservice for Destination-Bound Ride-Hailing Drivers ***Gülsüm Budakoglu*** *(Bitaksi)*	
133	09:10 – 09:25	#264 USA	Visual Editing and Debugging of Feature Models ***Cheng Thao*** *(Metropolitan State University), Ethan V. Munson (University of Wisconsin-Milwaukee)*	
134	09:25 – 09:40	#265 Türkiye	Issue Analyzer: Similarity-Based Grouping and Team Routing for Mobile Crash Logs *Selim Toksal (Trendyol Group),* ***Recep Çelik*** *(Trendyol Group), Berkay Mutlu (Trendyol Group), Tun Çakar (MEF University)*	
135	09:40 - 09:55	#269 Zambia	Design Science Research for Digital Platforms in Resource-Constrained Environments: The Lumikiza Case ***Morley Mujansi*** *(Copperbelt University), Alice P.S Shemi (Copperbelt University), Jeremiah J Mwiinga (Namibia University of Science and Technology), Lilian Mungulube (Zambian Open University)*	
136	09:55 – 10:00	#270 USA, Macedonia	Brain-Computer Interface, Brain Science, and Artificial Intelligence ***Stevo Bozinovski*** *(South Carolina State University), Adrijan Božinovski (American University of Europe FON)*	
137	10:00 – 10:15	#275 Türkiye	Scalability Analysis of CNN-Based sEMG Hand Gesture Recognition Across Increasing Gesture Repertoires ***Yasir ÖZLÜK*** *(Gazi University), Eda AKMAN AYDIN (Gazi University)*	
138	10:15 – 10:25	#281 Ukraine	SpecGuard: Multi-Layer Quality Assessment of Natural-Language FPGA Requirements ***Anton Striapunin*** (National Aerospace University "Kharkiv Aviation Institute)	
139	10:25 – 10:35	#208 South Africa	Evaluating Scene Localization Robustness in Mobile LiDAR AR Crime Scene Reconstruction under Geometric Degradation ***Kagiso Lawrence Pitsoane****, Mfundo Andrew Maneli, Omowunmi Elizabeth Isafiade (University of the Western Cape)*	
140	10:35 – 10:45	#203 Israel	Research Program: Computer-Aided Molecular Target Identification for Tuftsin-Phosphorylcholine in Autoimmune Diseases *Oleg Kupervasser (Shamoon College of Engineering), Marina Bershadsky (Shamoon College of Engineering), Gilad Halpert (Sheba Medical Center), Yehuda Shoenfeld (Sheba Medical Center),* ***Vadim Povolotskii*** *(Shamoon College of Engineering), Mariia Lishnevskii (Shamoon College of Engineering)*	
141	10:45 – 10:55	#282 Türkiye	Counterfactual Explanations for Gallstone Prediction: A Comparative Evaluation of DiCE and PertCF ***Busra Duygu Çelik*** *(Gazi University), Semiha Güngör (Hacı Bayram Veli University), Adem Tekerek (Gazi University)*	

142	10:55 – 11:05	#283 Türkiye	Classification of Eight Different Retinal Conditions in Retinal OCT Images Using Convolutional Neural Networks ***Semiha Güngör*** *(Hacı Bayram Veli University), Büşra Duygu Çelik (Gazi University), Adem Tekerek (Gazi University)*
143	11:05 – 11:15	#284 Türkiye	AI-Assisted Error Analysis to Design a Writing Course Syllabus ***Tuğba Dörterler*** *(Gazi University), Busra Duygu Çelik (Gazi University)*

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Innovative yet Low-Resource Uses of AI in Teaching: Pedagogical Boundaries and Ethical Considerations for Teachers

Mehriban A. Hasanova
Department of Computer Science and Algebra
Ganja State University, Ganja, Azerbaijan
Department of Automation and Information Technologies
University of Technology of Azerbaijan, Ganja, Azerbaijan
mehriban-70@mail.ru ; mehriban.hasanova@gdu.edu.az
ORCID: 0000-0002-7627-5196

Abstract
Generative artificial intelligence is increasingly entering educational practice. Yet, many current discussions still assume access to advanced infrastructure, paid platforms, and institutional support that teachers in low-resource settings often lack. This paper addresses that gap by proposing a conceptual and scenario-based framework for low-threshold AI use grounded in three criteria: feasibility in constrained contexts, preservation of teacher agency, and ethical low-risk application. Three scenarios are considered: AI-supported lesson planning, reflective practice, and task differentiation. Across them, AI is framed not as an instructional authority but as a bounded support tool for drafting, questioning, and structuring under teacher control. The paper argues that low-resource use of AI can provide a realistic, responsible, and professionally disciplined pathway for integrating generative AI into everyday teaching.

Keywords: artificial intelligence in education; low-resource teaching contexts; teacher-centered innovation; pedagogical boundaries; ethical use of AI; teaching practice

I. INTRODUCTION

Artificial intelligence (AI) has become an increasingly visible part of educational discourse, with growing expectations regarding its capacity to support teaching, personalize learning, and improve instructional decision-making [1], [2]. More recently, generative AI has intensified these discussions by expanding access to text-based support tools while also raising new concerns related to governance, implementation, and educational responsibility [3], [4]. Yet much of the current literature and many implementation models still assume access to robust digital infrastructure, subscription-based platforms, and sustained institutional support.

These assumptions reduce the relevance of many AI-in-education models for teachers working in low-resource contexts shaped by limited technology, restricted budgets, heavy workloads, and insufficient professional development. In such settings, the key issue is not whether AI can be used, but how it can be used meaningfully without undermining teacher judgment or increasing ethical risk. This paper, therefore, proposes a practice-oriented framework for the use of low-resource AI in teaching, focusing on feasible applications that remain pedagogically bounded and ethically responsible.

II. CONCEPTUAL AND DESIGN APPROACH

This study adopts a conceptual and scenario-based design approach. It is not presented as an empirical intervention, but as a practice-oriented framework for identifying realistic uses of generative AI in low-resource teaching contexts. The proposed scenarios were selected based on three criteria: feasibility in constrained settings, preservation of teacher agency, and ethical, low-risk use.

First, each scenario had to be implementable with freely available or open-access text-based AI tools, without requiring specialized software, advanced technical expertise, or data-intensive infrastructure [4]. Second, AI had to function only as a support tool for drafting, structuring, or reflection rather than as an instructional authority. Third, each scenario had to avoid using identifiable student data, automated profiling, and the delegation of assessment or instructional judgment to AI systems [3], [5]. Based on these criteria, three scenarios were selected because they correspond to recurring dimensions of everyday teaching practice: lesson planning, reflective professional thinking, and task differentiation for inclusive instruction.

III. INNOVATIVE LOW-RESOURCE USES OF AI IN TEACHING PRACTICE

Building on the conceptual and design-oriented framework outlined above, this section presents three worked scenarios that show how low-resource AI use can support everyday teaching practice while remaining pedagogically bounded and ethically responsible. Freely available, open-access generative AI tools offer opportunities for incremental innovation that can be integrated into existing teaching routines without requiring specialized infrastructure or extensive technical training [6].

1. AI-Supported Simplification of Lesson Planning

Lesson planning is a cognitively demanding and time-intensive component of teaching, particularly in low-resource contexts where teachers often work under conditions of high curricular pressure, limited preparation time, and restricted access to ready-made instructional materials. In such settings, the value of generative AI does not lie in replacing the pedagogical work of planning, but in supporting the early structuring of instruction through low-threshold, text-based assistance. Used appropriately, open-access AI tools can help teachers generate preliminary lesson outlines, propose sequences of classroom activities, and draft formative assessment prompts that can later be reviewed and refined through professional judgment.

The pedagogical significance of this support is especially apparent in subjects that require teachers to balance conceptual explanation, procedural modeling, classroom interaction, and differentiated learner engagement within a limited lesson period. One illustrative example is a Grade 8 Informatics lesson on algorithmic thinking and flowchart design. This topic requires more than the transmission of factual knowledge: it involves conceptual clarity, the sequencing of examples, opportunities for guided practice, and the translation of abstract ideas into manageable student activity. For a teacher working in a low-resource school environment, preparing such a lesson from scratch may require substantial time and cognitive effort, particularly when materials must be adapted without relying on specialized software or digital infrastructure.

In this scenario, a teacher may use a freely accessible generative AI tool to create an initial lesson structure by entering a prompt such as: “Create a 40-minute Grade 8 Informatics lesson outline on algorithmic thinking and flowchart design. Include a short warm-up based on everyday problem-solving, a teacher-guided explanation, one paired classroom task, and a brief formative assessment. The lesson should rely only on ordinary classroom resources and should support mixed-ability learners.” In response, the AI might generate a draft sequence that begins with an introductory discussion on everyday step-by-step processes, followed by an explanation of algorithms, sequence, and decision points, then a paired activity in which students convert a familiar everyday task into a simple flowchart, and finally a short exit task requiring students to identify errors in a partially completed flowchart.

At this stage, however, the generated output remains only a provisional scaffold. The pedagogical quality of the lesson still depends on the teacher’s interpretation and adaptation of that draft. The teacher may decide to simplify the terminology used during the explanation phase, replace generic examples with culturally or contextually familiar ones, adjust timing to the class's actual pace, and introduce additional scaffolding for students with limited prior exposure to symbolic representations. Similarly, for stronger learners, the teacher

may extend the paired activity by asking them to compare two possible flowchart structures for the same task and justify which one is clearer or more efficient. In this way, AI reduces the burden of initial lesson structuring without displacing the teacher's role as the designer of learning.

From a pedagogical standpoint, the contribution of AI in this context is therefore not automation, but assisted preparation. It can help teachers move more efficiently from instructional intention to a usable draft structure, especially where time, materials, and collegial support are limited. Yet the pedagogical boundary remains explicit: AI may support the drafting of a lesson outline, but it does not determine learning objectives, interpret learner misconceptions, select the final level of task difficulty, or make assessment decisions. Those responsibilities remain fully with the teacher. In low-resource contexts, this bounded use of AI is particularly valuable because it offers practical support while preserving contextual judgment, professional autonomy, and accountability for student learning

2. AI-Supported Reflective Teaching Practices

Reflective practice is widely recognized as a central dimension of effective teaching and professional growth. However, in low-resource educational contexts, sustained reflection is often constrained by heavy workloads, limited access to mentoring, and the absence of structured professional dialogue. Under such conditions, teachers may have valuable classroom insights but insufficient time or support to analyze them systematically. Generative AI can offer modest but meaningful assistance in this area by helping teachers structure reflective inquiry without displacing professional judgment.

The pedagogical value of such use lies in supporting reflection as a disciplined process rather than as an occasional impressionistic activity. In this scenario, AI is not used to evaluate teaching quality, assign performance judgments, or prescribe "correct" instructional decisions. Instead, it is used to expand the teacher's reflective field by generating questions, alternative interpretations, or lines of inquiry based on the teacher's own post-lesson observations. This bounded role is particularly suitable for low-resource settings because it requires only a short written input from the teacher and does not depend on classroom recordings, platform analytics, or student-level data.

This scenario may be illustrated by continuing the Grade 8 Informatics lesson on algorithmic thinking and flowchart design described in the previous section. After the lesson, the teacher writes a short reflective note such as the following: "Students were highly engaged during the opening discussion on everyday step-by-step processes, and most pairs were able to start the flowchart task. However, several students confused sequence and decision symbols, and the explanation stage took longer than expected. The final exit task showed that some students could identify flowchart errors, but others still seemed uncertain about when a decision point is needed." The teacher may then enter a prompt such as: "Based on this post-lesson note, generate five reflective questions to help a teacher analyze what worked well, what students found difficult, and what could be adjusted in the next lesson. Do not evaluate the teacher. Focus only on reflective inquiry."

In response, the AI might generate questions such as: Which part of the opening discussion appeared to make the concept of algorithm most accessible to students? At what point did confusion between sequence and decision symbols become most visible? How might the explanation stage be shortened while preserving conceptual clarity? What additional example or representation could help students distinguish between linear steps and decision points? How might the next lesson begin in a way that revisits uncertainty without discouraging learners who struggled? These outputs do not replace reflection; rather, they help the teacher move from a general impression of the lesson to a more structured pedagogical analysis.

The teacher remains fully responsible for interpreting these questions and deciding which insights are educationally meaningful. The AI does not know the classroom atmosphere, the prior attainment of particular learners, or the contextual reasons why some activities succeeded while others did not. Its contribution is

limited to supporting professional noticing and reflective questioning. In this sense, AI functions as a cognitive mirror rather than an evaluator: it reflects possible angles of analysis, but it does not determine conclusions. This distinction is essential. Reflection remains grounded in teacher-authored observations, contextual understanding, and professional values.

The ethical boundary in this scenario is equally important. No identifiable student data are entered into the system, and no student is individually described, classified, or profiled. The teacher's note remains focused on patterns of engagement, misunderstanding, pacing, and instructional flow rather than on personal learner records. Such use aligns with the principle of data minimization and reduces the risk of inappropriate disclosure. In low-resource contexts, where formal policy guidance may be limited, this low-input, low-risk form of reflective support is a realistic and responsible way to engage with AI.

From a pedagogical perspective, the contribution of AI here lies in strengthening reflective capacity under constrained conditions. Rather than offering automated advice, it helps teachers organize reflection in ways that may support more intentional follow-up planning, clearer recognition of learners' difficulties, and stronger continuity between lessons. The pedagogical boundary, however, remains explicit: AI may prompt reflective analysis, but it does not interpret teaching quality, determine pedagogical priorities, or replace the teacher's professional reasoning.

3. AI-Supported Task Differentiation for Inclusive Teaching

Task differentiation is an important but demanding aspect of inclusive teaching, particularly in classrooms where students vary in prior knowledge, pace of learning, language confidence, and readiness for abstraction. In low-resource settings, teachers are often expected to respond to this diversity without access to specialized support materials, adaptive software, or substantial preparation time. Under such conditions, generative AI can assist by producing alternative versions of the same task at different levels of complexity while preserving a common learning objective.

This form of support is especially relevant in subjects such as Informatics, where conceptual understanding and procedural reasoning may develop unevenly across learners. Continuing the Grade 8 lesson on algorithmic thinking and flowchart design, a teacher may wish to prepare differentiated follow-up tasks after noticing that some students are still uncertain about basic symbol use. In contrast, others are ready to analyze more complex structures. Rather than designing each variation from the beginning, the teacher may use a prompt such as: "Create three versions of a classroom task on flowchart design for Grade 8 students: one simplified version for learners who need more structure, one standard version, and one extension task for learners ready for deeper reasoning. All three tasks should address the same learning objective: understanding sequence and decision points in a simple algorithm. Do not require computers or specialized software."

In response, the AI might generate a simplified task in which students complete a partially structured flowchart using a limited set of symbols and labelled prompts; a standard task in which students design a full flowchart for an everyday process such as borrowing a library book or preparing for school; and an extension task in which students compare two different flowchart solutions for the same problem and explain which one is logically clearer or more efficient. These task variations share a common conceptual focus but differ in the degree of independence, abstraction, and explanation required.

The instructional value of this support lies in accelerating the production of parallel learning tasks without fragmenting the lesson into unrelated activities. The teacher can review the generated options, adjust wording, simplify or enrich the demands of each version, and ensure that the final tasks remain aligned with the curriculum and appropriate for the actual classroom context. This is particularly useful in constrained environments, where differentiation is often recognized as pedagogically necessary but is difficult to implement consistently due to time pressure.

At the same time, the educational boundary in this scenario is specific and important: AI may assist in generating alternative task formats, but it does not decide which learner receives which task, nor does it infer ability profiles from student data. Decisions about grouping, task allocation, pacing, and support remain matters of teacher judgment. This distinction helps preserve inclusive practice as a pedagogical process rather than allowing it to drift into automated categorization. In this respect, low-resource use of AI is most defensible when it supports preparation for differentiated teaching without automating decisions about learners.

Table 1. Mapping low-resource AI use scenarios to pedagogical boundaries and ethical safeguards

Scenario	Primary teaching needs	Example AI support	Pedagogical boundary	Ethical safeguard
AI-supported lesson planning	Reducing the time and cognitive load involved in structuring lessons under constrained conditions	Drafting a preliminary lesson outline, suggesting a sequence of activities, and proposing brief formative assessment prompts	AI supports initial structuring only; final decisions on objectives, sequencing, explanation, and assessment remain with the teacher.	No identifiable student data are required; AI-generated plans are treated as provisional drafts subject to teacher review and contextual adaptation.
AI-supported reflective practice	Supporting structured post-lesson reflection where time, mentoring, or professional dialogue may be limited	Generating reflective questions or alternative angles of analysis from teacher-authored post-lesson notes	AI does not evaluate teaching quality or prescribe pedagogical judgments; reflection remains grounded in the teacher's own interpretation.	Reflection is based on non-identifiable classroom observations; no personal learner data, profiling, or individual records are entered.
AI-supported task differentiation	Preparing parallel versions of learning tasks to support inclusive teaching across varied learner readiness levels	Producing simplified, standard, and extension versions of tasks aligned with the same learning objective	AI may suggest differentiated task formats, but task allocation, grouping, and instructional adjustment remain teacher decisions.	No automated grouping or learner profiling is used; differentiation remains a pedagogical strategy rather than a technical classification process.

As Table 1 indicates, the educational value of low-resource AI use lies not in autonomous instructional decision-making, but in bounded forms of support that remain pedagogically guided and ethically constrained.

IV. PEDAGOGICAL BOUNDARIES OF AI USE IN TEACHING

The educational value of low-resource AI use depends not on the expansion of automation, but on the clarity of its pedagogical limits. In the scenarios outlined above, AI contributes most meaningfully when it supports preparatory and reflective aspects of teaching without assuming instructional authority. For this reason, responsible AI use in teaching requires a bounded framework in which pedagogical purpose remains primary and professional judgment remains non-transferable.

The first boundary concerns the location of instructional authority. AI may assist with drafting lesson structures, generating reflective prompts, or proposing differentiated task formats. Still, it does not determine learning objectives, interpret classroom realities, or decide how instruction should be adapted for particular learners. These functions depend on contextual judgment, curricular awareness, and knowledge of students that remain irreducibly human and professionally situated. AI output is therefore pedagogically useful only when treated as provisional material for teacher review rather than as guidance to be implemented uncritically.

A second boundary concerns evaluative responsibility. Even where AI can help organize ideas, suggest feedback language, or surface possible misconceptions, it must not displace the teacher's role in assessment, interpretation, and follow-up decision-making. Educational judgment is not merely a technical act of classification; it is a relational and normative process tied to fairness, context, and professional accountability. In this respect, bounded use of AI protects not only instructional quality but also the integrity of assessment, a core pedagogical responsibility.

A third boundary concerns the relationship between efficiency and educational purpose. In low-resource settings, the appeal of AI often lies in its capacity to save time or reduce workload. While such efficiency may be valuable, it is not in itself a sufficient pedagogical justification. AI use becomes educationally defensible only when it strengthens instructional coherence, supports inclusion, deepens teacher reflection, or otherwise contributes to meaningful learning processes. The criterion is therefore not whether AI can do something quickly, but whether its use serves a recognizable educational purpose under teacher direction.

Taken together, these boundaries position AI not as a teaching agent, but as a constrained pedagogical support tool. Such a position is especially important in low-resource contexts, where the absence of extensive infrastructure or institutional regulation can increase the temptation to use AI pragmatically without sufficient educational scrutiny. Clear pedagogical limits help ensure that innovation remains teacher-centered, context-sensitive, and aligned with the core purposes of teaching. This bounded approach also provides the necessary foundation for the ethical considerations discussed in the following section.

V. ETHICAL CONSIDERATIONS IN LOW-RESOURCE AI USE

Ethical questions surrounding artificial intelligence in education become particularly significant in low-resource contexts, where institutional guidance, technical safeguards, and digital literacy may be uneven or underdeveloped. Under such conditions, responsible AI use cannot rely solely on formal regulation or platform design; it must also be sustained through clear professional norms. For this reason, ethical considerations are not secondary to pedagogical concerns but are an essential condition for the adoption of low-resource AI in teaching [3], [5].

A first ethical priority is data minimization. Many AI systems operate through large-scale data processing, which raises concerns about what information is entered, how it may be stored, and whether it could later be reused in ways that are not fully visible to teachers. In low-resource classroom practice, the most defensible approach is therefore to avoid entering identifiable student information altogether. The scenarios proposed in this paper are intentionally designed around teacher-authored notes, generic lesson content, and non-identifiable task materials rather than individual learner records. This reduces privacy risk while keeping AI use compatible with ordinary classroom realities [3].

A second ethical consideration is transparency of use. When AI contributes to lesson preparation, reflective questioning, or task design, that support should not be concealed behind an appearance of purely teacher-generated authority. Transparency matters not because every technical detail must be disclosed, but because teachers and institutions should maintain clarity about the role AI has played in shaping instructional materials. In low-resource settings, where misconceptions about AI may be widespread, such openness helps preserve trust. It prevents the mistaken assumption that AI-generated outputs are inherently objective, authoritative, or pedagogically neutral [3], [5].

A third issue concerns the avoidance of learner profiling and automated categorization. This is especially important in contexts where educational disadvantage may already be shaped by language background, unequal access to resources, or variable prior preparation. If AI is used to classify students too quickly into ability-based categories, it risks reinforcing existing inequities under the appearance of technical efficiency. For this reason, AI may support differentiated task design, but the interpretation of learner needs and the allocation of tasks must remain pedagogical decisions rather than algorithmic ones. Ethical low-resource AI use therefore resists the drift from instructional support toward hidden classification [5].

A fourth ethical concern is the preservation of clear responsibility. Even where AI contributes to drafting, structuring, or idea generation, responsibility for instructional choices and their consequences remains human and professional. In low-resource environments, where policy frameworks may lag behind practice, this clarity is particularly important. Ethical use of AI requires that teachers remain able to explain, justify, and, if necessary, revise the educational decisions associated with AI-supported materials. The ethical issue here is not only who uses AI, but who remains answerable for its pedagogical consequences [3].

Taken together, these considerations suggest that the ethical legitimacy of low-resource AI use lies less in the technology's sophistication than in the discipline of its application. Simplicity, restraint, and contextual awareness are not signs of limited innovation; they are conditions for responsible adoption. In this sense, ethical low-resource AI use is best understood as a cautious professional practice that enables teachers to benefit from generative tools without normalizing opacity, overreach, or new forms of inequity.

VI. IMPLICATIONS FOR TEACHING PRACTICE

The analysis developed in this paper has implications not only for individual classroom practice but also for teacher learning and institutional support in low-resource educational settings. Its central practical implication is that meaningful AI integration does not require comprehensive technological transformation. Instead, it can begin through bounded, low-threshold uses that address familiar teaching tasks while remaining compatible with existing classroom realities.

For teachers, this means that AI adoption may be approached as a form of disciplined pedagogical support rather than as a demand for technical specialization. Low-resource uses such as lesson structuring, reflective prompting, and task variation offer manageable entry points because they do not depend on advanced infrastructure, large datasets, or formal automation. Their practical value lies in helping teachers work more efficiently in areas where preparation time, material support, or collegial collaboration may be limited.

For teacher education and professional development, the findings suggest that AI-related preparation should focus less on mastery of specific platforms and more on cultivating transferable judgment. Pre-service and in-service programs can use low-resource scenarios to help teachers examine when AI support is educationally useful, where its limits should be drawn, and how ethical considerations shape classroom decision-making. In this sense, AI competence in teaching should be understood not primarily as technical fluency, but as the ability to use generative tools selectively, critically, and contextually [7].

At the institutional level, the paper points toward the need for flexible guidance rather than uniform mandates. Schools and teacher education providers should recognize that educational contexts differ significantly in infrastructure, policy readiness, and staff confidence. In settings where formal AI frameworks are still emerging, a minimal responsible-use protocol may be more realistic and more useful than ambitious implementation models [8], [9]. Such a protocol may include avoiding the use of identifiable student data, requiring teacher review of all AI-supported materials, maintaining transparency in use, and prohibiting automated learner classification or assessment decisions.

Taken together, these implications suggest that sustainable AI integration in education is more likely to develop through professionally governed, context-sensitive practices than through top-down technological expectation. In low-resource settings, especially, the most durable forms of innovation may be those that remain modest in technical demands but clear in pedagogical purpose [9].

VII. CONCLUSION

This paper has argued that generative AI can have educational value in low-resource contexts when its use is framed not by technological ambition, but by pedagogical discipline. Rather than treating limited infrastructure, time, or institutional support as conditions that exclude meaningful innovation, the study has shown that such contexts may instead call for more selective, bounded, and teacher-centered forms of AI adoption.

By organizing the discussion around lesson planning, reflective practice, and task differentiation, the paper has identified three areas in which low-threshold AI use can support everyday teaching without displacing professional judgment. Its central contribution lies in showing that the question is not simply whether AI is available, but under what pedagogical and ethical conditions its use becomes educationally defensible.

The framework proposed here, therefore, shifts attention away from high-resource models of AI integration and toward a more realistic account of responsible classroom use. In this account, accessibility, contextual adaptation, and professional restraint are not signs of reduced innovation; they are part of what makes the use of AI sustainable and educationally credible. This perspective may be especially relevant for systems in which teachers are expected to engage with AI before comprehensive policy, infrastructure, or technical support is in place [3], [4]. As generative AI continues to develop, its educational significance will depend less on its novelty than on the quality of the professional boundaries within which it is used.

This paper is intentionally conceptual and scenario-based and therefore does not claim empirical validation of the proposed framework. Its purpose is to clarify how low-resource uses of generative AI may be structured in pedagogically bounded and ethically responsible ways, rather than to test their effectiveness through classroom intervention. The three scenarios presented here are intended as illustrative models of practice, designed to support discussion, adaptation, and future application. Further research is needed to examine these uses empirically across subjects and school settings, particularly regarding teacher uptake, contextual adaptation, perceived usefulness, and the institutional conditions required for sustainable and responsible implementation.

REFERENCES

[1] W. Holmes, M. Bialik, and C. Fadel, Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign, 2019.

[2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education – Where are the educators?” International Journal of Educational Technology in Higher Education, vol. 16, Art. no. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[3] F. Miao and W. Holmes, Guidance for generative AI in education and research. UNESCO, 2023.

[4] OECD, OECD Digital Education Outlook 2026: Exploring effective uses of generative AI in education. OECD Publishing, 2026, doi: 10.1787/062a7394-en.
[5] I. M. García-López and L. Trujillo-Liñán, "Ethical and regulatory challenges of generative AI in education: A systematic review," Frontiers in Education, vol. 10, Art. no. 1565938, 2025, doi: 10.3389/feduc. 2025.1565938.
[6] A. Alfarwan, "Generative AI use in K-12 education: A systematic review," Frontiers in Education, vol. 10, Art. no. 1647573, 2025, doi: 10.3389/feduc. 2025.1647573.
[7] X. Tan, K. S. Cheng, and M. H. A. Ling, "Artificial intelligence in teaching and teacher professional development: A systematic review," Computers and Education: Artificial Intelligence, vol. 8, Art. no. 100355, 2025, doi: 10.1016/j.caeai.2024.100355.
[8] Y. H. Cheah, J. Lu, and J. Kim, "Integrating generative artificial intelligence in K-12 education: Examining teachers' preparedness, practices, and barriers," Computers and Education: Artificial Intelligence, vol. 8, Art. no. 100363, 2025, doi: 10.1016/j.caeai.2025.100363.
[9] D. T. K. Ng, E. K. C. Chan, and C. K. Lo, "Opportunities, challenges and school strategies for integrating generative AI in education," Computers and Education: Artificial Intelligence, vol. 8, Art. no. 100373, 2025, doi: 10.1016/j.caeai.2025.100373.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Methodological Framework for Preparing Future Computer Science Teachers for Sustainable Professional Activity in the Artificial Intelligence Context

Aizhan Adambekova
Department of Informatics and Digitalization of Education
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
aizheke@gmail.com
ORCID: 0009-0009-3200-2194

Aigul Aldabergenova
Department of Informatics and Digitalization of Education
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
a.aldabergenova@zu.edu.kz
ORCID: 0000-0003-2404-7615

Aiym Yesseikyzy
Department of Informatics and Digitalization of Education
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
aiym.yesseykyzyy@gmail.com
ORCID: 0000-0002-1216-5828

Ulzhalgas Yesseikyzy
Department of Informatics and Digitalization of Education
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
yesseikyzy@gmail.com
ORCID: 0000-0002-8458-743X

Gulmira Kasteyeva
Department of Informatics and Digitalization of Education
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
kasteevagd@gmail.com
ORCID: 0000-0001-6536-6933

Abstract
The integration of artificial intelligence in education demands new approaches to computer science teacher preparation that go beyond technical training. This paper presents a three-component model treating technical, ethical, and social competencies as coequal, interdependent pillars within a structural-functional framework. Its distinguishing feature—compared to existing frameworks such as DigCompEdu, TPACK, and AI4K12, which treat ethics as a sub-competency or supplementary module—is the cross-cutting integration matrix (3×4), which operationalizes all three dimensions simultaneously across every component of professional readiness. The model is grounded in SDG 4 and informed by a survey of 899 CS teachers across five regions of Kazakhstan ($\alpha = 0.93–0.97$), revealing critical gaps in AI competency ($M = 3.48$). Six pedagogical conditions for implementation and practical implications for curriculum developers are discussed.

Keywords: artificial intelligence, AI ethics, computer science teacher preparation, SDG 4, three-component model, DigCompEdu, structural-functional approach.

I. INTRODUCTION

Artificial intelligence has moved far beyond the boundaries of computer science and is now shaping economies, healthcare systems, public administration, and – most importantly – education. By 2025, AI tools had reached massive scale: ChatGPT surpassed 700 million weekly active users, fundamentally transforming learning, teaching, and the way educational institutions function [1]. Against this backdrop, the global sustainable development agenda takes on new meaning: the UN Sustainable Development Goals adopted in 2015, and SDG 4 in particular – ensuring quality, inclusive, and equitable education for all [2] – now face a fundamentally new challenge. These two forces – AI and sustainable development – converge on one of the most pressing questions of our time: how to prepare teachers who can convey not only technical knowledge but also ethical responsibility and the social dimension of technology. Existing models of computer science teacher preparation primarily focus

on technical and methodological training. An analysis of international frameworks shows that the ethical dimension appears only as a sub-competency (DigCompEdu) or is absent as a central focus altogether (TPACK), and the social dimension is only partially addressed [3], [4]. UNESCO explicitly calls for integrating AI ethics into teacher education [2], and the OECD has documented a critical gap between teachers' technical skills and their readiness to analyze technological systems [5]. This gap is particularly relevant in the context of Kazakhstan's national policy, which has committed to digitalization and the development of AI literacy. Kazakhstan's Education Development Concept for 2023–2029 provides for updating computer science content, introducing AI elements into curricula, and training a new generation of teachers [6]. However, structural mechanisms to integrate ethical and social competencies into this preparation remain underdeveloped. This paper presents an AI-integrated three-component model for computer science teacher preparation, grounded at the intersection of international competency frameworks (DigCompEdu, TPACK, AI4K12), the Sustainable Development Goal (SDG) 4, the national context of Kazakhstan, and empirical data from a large-scale teacher survey. The scientific novelty lies in three contributions absent from prior work: (1) elevating ethical and social competencies to structural, non-optional elements of every training component rather than separate electives; (2) operationalising this integration through a cross-cutting matrix (3×4) that functions as both a curriculum design tool and an audit instrument; and (3) contextualising global frameworks (DigCompEdu, TPACK, AI4K12) within a post-Soviet pedagogical tradition, thereby creating a model directly applicable to Kazakhstan and similar national contexts.

II. THEORETICAL BACKGROUND

1. International Competency Frameworks

The DigCompEdu framework [4] defines seven areas of digital competence for educators and six proficiency levels (A1–C2), emphasizing the need for critical reflection on technology. Area 6, "Empowering Learners' Digital Competence," includes elements of digital citizenship, but the ethical dimension of AI – bias analysis, algorithm auditing, and data protection is not identified as a standalone competency. The TPACK model [3] holds that effective technology integration requires the simultaneous operation of technological (TK), pedagogical (PK), and content (CK) knowledge at their intersections: a computer science teacher must not only understand how AI works but also know how to teach it. Yet neither TK nor TPK contains an explicit ethical or social component.

The AI4K12 framework [7] identifies five big ideas in AI for K-12 education (perception, representation, learning, interaction, and societal impact), with the fifth idea directly addressing social consequences. The UNESCO Recommendation on the Ethics of Artificial Intelligence [8] articulates principles of fairness, transparency, and accountability. However, none of these frameworks offers a pedagogical architecture that structurally integrates the ethical and social dimensions into all components of teacher preparation.

2. The Structural-Functional Approach: Rationale for Selection

Educational theory offers several frameworks that describe the cyclical nature of professional development. Kolb's experiential learning cycle [9] models learning through experience as a four-phase sequence (concrete experience → reflective observation → abstract conceptualization → active experimentation) but does not differentiate types of professional readiness. Schön's reflective practice [10] emphasizes the role of "reflection-in-action" but focuses on a single component. Shulman's pedagogical content knowledge (PCK) [11], which became the foundation of TPACK, describes the intersection of subject-matter and pedagogical knowledge but does not include a motivational component.

Three advantages justify the choice of the structural-functional approach. First, structural completeness: the approach explicitly identifies four qualitatively distinct components of readiness – motivational, cognitive, activity-based, and reflective. This creates an architecture that embeds three-component integration across all components simultaneously. Second, a built-in motivational component: professional identity is treated as a foundational element, and developing a conscious attitude toward the ethical and social dimensions of AI is a necessary precondition for lasting competence [12]. Third, contextual compatibility: the approach has a strong tradition in pedagogical scholarship in Kazakhstan and the post-Soviet region, ensuring the model is practically applicable within the national context. The structural-functional approach was thus chosen not as an alternative to Western models but as the most appropriate architectural foundation, synthesizing insights from different pedagogical traditions. Kolb's cycle informs the activity-based and reflective components; Shulman's PCK resonates with the cognitive. The proposed framework provides a structure for simultaneously developing all four components of readiness across all three dimensions.

III. EMPIRICAL PROBLEM STATEMENT

To assess the scope of the problem in the national context, a survey of 899 computer science teachers from five regions of Kazakhstan was conducted in July–August 2025. A mixed methodology was used: a quantitative component (Likert scale, 1–5) supplemented by open-ended questions. The sample was purposive, ensuring representativeness by school type (urban/rural), teaching experience, and qualification category. Instrument reliability: Cronbach's $\alpha = 0.93–0.97$ across scales. Results are presented in Table 1.

Table 1. Assessment Results for Computer Science Teacher Competencies

Competency Area	M	SD	Level
Office Applications	4.26	0.94	High
Cloud Services	4.02	0.93	High
Cybersecurity	3.73	0.95	Moderate
Algorithmic Thinking	3.63	0.84	Moderate
Artificial Intelligence	3.48	0.91	Below average

The results reveal a consistent pattern: teachers are confident with familiar tools (office applications: $M = 4.26$; cloud services: $M = 4.02$) but lag significantly in areas that define new directions in computer science — artificial intelligence ($M = 3.48$), algorithmic thinking ($M = 3.63$), and cybersecurity ($M = 3.73$). The spread of SD values indicates considerable variability in preparation: some teachers have already adopted new technologies, while a large share remains at a basic level. A regional analysis found statistically significant differences between urban and rural schools: urban teachers scored higher on AI competencies ($M = 3.72$ vs $M = 3.21$, $p < 0.01$, $\eta^2 = 0.013$), confirming that the gap is real. However, its modest effect size suggests it cannot be resolved by targeting individual teachers alone — structural support in preparation programs is required. A parallel deficit appears in methodological competencies ($\alpha = 0.97$; composite $M = 3.77$): while teachers rated online platform use highly ($M = 3.89$), AI use in teaching scored only $M = 3.66$ and project-based learning $M = 3.47$ — confirming that the gap is not merely technical but extends to the pedagogical application of new technologies. Teachers know the tools; they lack a framework for integrating them meaningfully into instruction. This is precisely the gap the proposed model is designed to address. Open-ended responses corroborated the quantitative findings: 68% of respondents noted a lack of specialized professional development courses on AI, and 43% reported a shortage of teaching materials in Kazakh and Russian.

It is important to note that these data represent the lower bound of the problem. The survey measured technical and methodological competencies but did not include questions about ethical competency (the ability to analyze algorithmic bias and conduct ethical audits of AI systems) or social competency (awareness of digital inequality and assessment of the social consequences of technology). If technical competency — the most measurable and traditionally cultivated dimension — already shows gaps, the ethical and social dimensions, which are not systematically developed in current training programs, are likely at an even lower level. International evidence supports this conclusion: UNESCO [2] and the OECD [5] document the systematic absence of ethical competencies in computer science teacher preparation programs; Nabi et al. [13] note that issues of algorithmic bias, ethics, and data privacy remain underrepresented in such programs.

Moreover, the technical deficit cannot be addressed in isolation. A teacher who masters the inner workings of a neural network but does not understand why it can discriminate is not ready to teach AI in school. Technical knowledge without an ethical context is incomplete knowledge. This defines the structural priorities of the proposed model: not simply strengthening technical training, but simultaneously developing technical, ethical, and social competencies as interdependent dimensions.

IV. THE THREE-COMPONENT INTEGRATION MODEL

The model operates on two levels. The first is the structural-functional framework, which describes the overall architecture of preparation through four cyclically linked components. The second – and this is where the main scientific contribution lies – is the three-component integration of technical, ethical, and social competencies, which constitutes the model's innovative core.

1. The Structural-Functional Framework

Table 2. Structural-Functional Components of the Model

Component	Function	Key Mechanisms
Motivational	Builds professional identity	Career guidance, mentoring, success experiences, and role models
Cognitive	Develops integrated knowledge	Three-component integration; AI literacy; pedagogical competence
Activity-based	Applies knowledge in practice	Project-based learning, teaching practice, hackathons, service-learning
Reflective	Consolidates through metacognition	Reflective journals, peer-feedback, video reflection, mentoring cycles

The four components operate cyclically: motivational readiness creates conditions for learning; cognitive development provides a foundation of integrated knowledge; activity-based learning deepens knowledge through practice; and reflection transforms experience into lasting competence and generates new motivational impulses. The cycles are not strictly sequential – all four components are active at every stage of the preparation program, but with varying intensity.

The motivational component plays a particularly important role in the context of three-component integration. Technical motivation is built through career orientation in the AI industry and the experience of successfully creating working systems. However, ethical and social motivation require different mechanisms: mentoring by practitioners who have faced real ethical dilemmas; analysis of cases in which technologies amplified inequality;

and developing awareness of the computer science teacher's social mission as a guide to responsible technology. Without a motivational foundation, students treat the ethical and social components as a formality rather than as an integral part of the profession.

2. Three-Component Integration

The model's innovative core treats three competencies as equally important, interdependent dimensions. Technical competency includes programming and algorithmic thinking, AI system architecture (machine learning, neural networks, generative AI), cybersecurity, and Green IT — awareness of the environmental impact of digital technologies, energy-efficient software design, and responsible resource consumption. Green IT is deliberately placed within the technical dimension rather than the ethical one because its content is primarily about engineering choices (algorithm efficiency, server energy consumption, hardware lifecycle management) rather than value judgments. A future teacher who can compare the carbon footprint of two equivalent algorithms is better equipped to address SDG 13 (Climate Action) in a classroom context, reinforcing the model's grounding in the sustainability agenda.

Ethical competency is defined as an individual teacher's capacity to evaluate the moral dimensions of a specific technological artifact or pedagogical decision: it covers algorithmic bias analysis (bias audit), the principles of transparency, explainability, and accountability articulated in the UNESCO Recommendation [8], personal data protection and privacy, and digital citizenship. Social competency, by contrast, operates at the level of structural and societal consequences: it encompasses awareness of digital inequality at both the global and national levels (e.g., the documented gap between urban and rural schools in Kazakhstan), analysis of AI's impact on the labor market, service-learning as a method for integrating social design into the curriculum, and Social Impact Assessment of technological decisions. The key distinction is analytical: ethical competency asks, "Is this particular system fair and responsible?"; social competency asks, "What broader inequalities does this technology reinforce or alleviate?" Both are necessary because a teacher can uphold data-privacy principles while remaining unaware that the same tool widens the digital divide.

These three dimensions are inseparable. Technical competency without an ethical foundation produces specialists who can build powerful but potentially dangerous systems – a neural network trained on biased data will reproduce discrimination regardless of its technical quality. Ethical thinking without technical literacy produces critics who cannot engage with the systems they evaluate – it is impossible to analyze algorithmic bias without understanding how the algorithm works. Social awareness without the other two produces activists disconnected from the mechanisms of change.

Three-component integration directly addresses SDG 4. The ethical component ensures the fairness of educational technologies through systematic bias analysis and the protection of student data. The social component ensures inclusivity by raising awareness of and taking action on digital inequality. The technical component ensures quality education through sound technology management. The overall architecture of the model is presented in Figure 1.

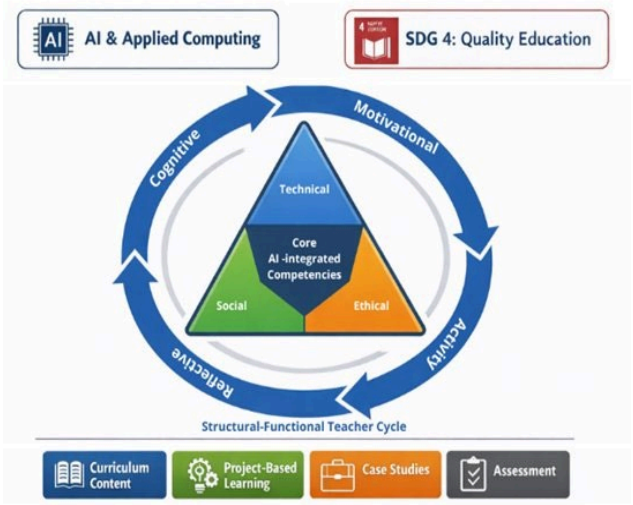


Figure 1. AI-integrated three-component model for computer science teacher preparation

3. Integration Levels and Implementation Examples

Integration is implemented at four levels. At the curriculum level, modules on AI ethics and social implications are embedded directly into technical courses rather than offered as separate electives.

At the case-study level, complex real-world scenarios are examined from all three perspectives. For example, the case "Using spreadsheets and databases in an informatics lesson" requires students to conduct a technical analysis — examining how data structures are organized and how sorting, filtering, and information retrieval are implemented; an ethical audit — ensuring responsible handling of student data without violating privacy or misinterpreting results; and a social evaluation — assessing whether the task creates additional barriers for students with different levels of digital competence and how it can be adapted for inclusive learning. At the same time, the discussion may incorporate elements of modern AI systems that operate on data, thereby expanding the context without going beyond the official curriculum.

Another example is the case "Using Python programming in student projects (e.g., developing a 2D game or a simple application)". Students analyze the technical aspects of development — program structure, algorithms, and interaction logic; ethical considerations — including acceptable content of digital products, responsibility for created materials, and issues of academic integrity when using pre-existing code; and social implications — such as differences in learning pace, access to devices, and availability of support outside school. Additionally, the impact of modern AI tools (e.g., code generators) on the learning process is examined, particularly how their use may either support or undermine students' independence.

Table 3. Sample Learning Objectives by Topic — "Methods of Teaching Computer Science" Course

Course Topic	Technical Objective	Ethical Objective	Social Objective
Designing a lesson on algorithms and data sorting	The student develops a lesson plan with clearly defined learning outcomes,	The student incorporates a case discussion into the lesson plan: what ethical	The student provides two differentiated versions of the lesson task — one for a

Course Topic	Technical Objective	Ethical Objective	Social Objective
	a step-by-step task structure, and assessment criteria aligned with the school curriculum.	risks arise when a sorting algorithm is applied to personal data (e.g., student records), and how should a teacher address this with pupils.	well-equipped urban school and one for a rural school with limited access to devices — and justifies the adaptations.
Designing a lesson on programming in Python	The student produces a full technological map of the lesson, including learning objectives, instructional methods, time allocation, and sample code tasks at three difficulty levels.	The student designs a classroom discussion protocol on academic integrity when pupils use AI code generators, specifying discussion questions and teacher response strategies.	The student analyses how unequal home internet access affects pupils' ability to complete programming homework and proposes at least one compensatory measure within the lesson design.

Table 3 illustrates how the three-component integration operates in a university methods course. Each objective describes what the student-teacher learns to design, analyze, and justify.

At the assessment level, evaluation criteria explicitly capture competencies across all three dimensions, making integration visible and accountable: the final project is assessed not only in terms of the technical quality of the developed instructional materials, but also in terms of the depth of ethical analysis of pedagogical decisions and the validity of the social evaluation of their impact on a specific group of learners.

4. The Cross-Cutting Integration Matrix

The key instrument of the model is the cross-cutting integration matrix (3×4), which shows how each of the three competency dimensions manifests across the four structural components of readiness (Table 4).

Table 4. Cross-Cutting Integration Matrix (3×4)

Dimension	Motivational	Cognitive	Activity-based	Reflective
Technical	Career orientation in AI	AI architecture, programming, cybersecurity	AI project development, hackathons	Technical analysis of decisions
Ethical	Professional responsibility of the teacher	AI ethics, bias analysis, data protection	Ethical audit of projects	Reflection on ethical dilemmas
Social	Awareness of social mission	AI and society, digital inequality	Service-learning, social design	Social impact analysis

The matrix illustrates the principle of structural simultaneity: no technical action is taken without ethical reflection and social evaluation. Each cell represents a specific educational objective that can be operationalized in a curriculum. For example, the cell "Ethical × Activity-based" is set to require a mandatory ethical audit for every student AI project, using a structured checklist. This distinguishes the model from frameworks that treat ethics as an add-on rather than a foundational element.

To make the matrix actionable for curriculum developers, a sample assessment rubric for a capstone AI lesson-design project assigns equal weight to each dimension: (a) Technical — correct AI tool selection, algorithmic soundness, cybersecurity safeguards in place (33%); (b) Ethical — bias audit completed with reference to UNESCO ICT-CFT principles, data-privacy protections explicitly documented (33%); (c) Social — explicit adaptation for rural or low-resource learners, inclusivity analysis provided (33%). The equal weighting is itself a pedagogical statement: it prevents the technical dimension from crowding out the other two, a documented risk when rubric design is not aligned with the integration model [12].

5. Pedagogical Conditions for Implementation

A review of the literature and the barriers identified in the empirical study led to the identification of six pedagogical conditions necessary for successful implementation of the model:

1. Integration of AI ethics into all computer science modules – not as a separate course but as a thread running through every discipline [2], [8]. In practice, this means a machine learning lab assignment should include a task on analyzing bias in the training dataset.

2. Availability of AI simulation tools and a case study library for modeling ethical dilemmas in a safe environment.

3. Interdisciplinary collaboration between departments of computer science and the humanities (philosophy, sociology), ensuring expertise in ethics and social theory.

4. Ongoing professional development for university faculty in AI ethics – a teacher who has no personal experience with ethical analysis of technology cannot develop that competency in students.

5. An institutional culture that values ethical reflection on a par with technical achievement – without this, students will treat the ethical component as a formality with no bearing on academic success.

6. Alignment of assessment criteria with all three competency dimensions: if 90% of the marks in a capstone project rubric go to technical execution and only 10% to ethical analysis, the integration model remains declarative.

Without all of these conditions being met simultaneously, a model that integrates ethical and social dimensions in theory may, in practice, default to a primarily technical program.

V. DISCUSSION

The proposed model makes three contributions to the existing literature. First, it elevates the ethical and social dimensions from supplementary modules to structural elements integrated into all components of readiness. This is a key distinction from approaches in which “AI ethics” is offered as a separate elective course, isolated from technical training. Research on computer science teacher preparation shows that lasting professional competencies develop when key content elements are embedded in disciplinary and pedagogical training rather than studied outside that context [12]. Analogous to the integration of computational thinking into teacher preparation programs, the ethical and social dimensions of AI should be embedded directly into the structure of technical disciplines, ensuring their organic integration into future teachers’ professional identity.

Second, the 3×4 matrix offers a concrete curriculum design tool. Unlike general declarations about the need for ethical education, the matrix shows exactly where each dimension appears at each stage of preparation. This shifts the discussion from declarations to design: a curriculum developer can use the matrix as a checklist to verify that each cell contains specific educational content.

Third, the model was developed at the intersection of international frameworks and Kazakhstan's national context, demonstrating how global principles can be adapted to regional conditions. The structural-functional approach, with its strong tradition in post-Soviet pedagogy, serves as a "translator" between international competency frameworks and local teacher preparation practice.

For Kazakhstan, the model is particularly relevant in light of the implementation of the Education Development Concept for 2023–2029 [6], which provides for updating computer science content and introducing AI elements. The current Concept emphasizes the technical aspects of digitalization. Still, the proposed model demonstrates that sustainable digital transformation of education is impossible without the parallel development of teachers' ethical and social competencies. The 3×4 matrix can serve as an audit tool for existing computer science teacher preparation programs at Kazakhstani universities, identifying "empty cells" – areas where the ethical or social dimension is absent.

The model complements rather than replaces existing frameworks. TPACK remains relevant for describing the intersection of knowledge types; DigCompEdu for assessing levels of digital competence; AI4K12 for content-level guidance. The proposed model adds an architectural foundation that ensures the structural integration of ethical and social dimensions across all components of professional readiness, functioning as a "meta-framework."

The study has several limitations. First, the model at this stage is a theoretical construct informed by empirical data but not yet experimentally validated. The empirical survey established the problem's relevance, not the model's validity. Second, reliance on self-reported data may overestimate actual competency levels: teachers tend to rate their own skills higher than objective tests would show. Third, the purposive sample, limited to five regions of Kazakhstan, does not allow generalization to the entire national education system. Fourth, as noted above, the survey instrument did not directly measure ethical and social competencies, a deliberate methodological limitation rather than an oversight.

At the same time, the very absence of validated instruments to measure the ethical and social competence of computer science teachers underscores the innovative nature of the proposed model and points the way for future work. How might ethical and social competencies be measured in practice? Existing psychometric approaches offer a useful starting point. Ethical competency can be assessed using scenario-based instruments, in which respondents analyze a vignette describing a biased recommendation algorithm or a data privacy breach. They are scored on the accuracy of their bias identification, the quality of their remediation proposal, and their reference to relevant principles. Such instruments echo established approaches in medical ethics education (Objective Structured Clinical Examination) and can be adapted to teacher preparation contexts. Social competency, operating at a structural level, is better captured through portfolio and performance assessments: for example, a Social Impact Assessment rubric applied to a student-designed lesson plan, or a structured reflection on how a specific tool might differentially affect urban and rural learners. Pre/post quasi-experimental designs using these instruments, combined with a Sustainable Digital Competency diagnostic instrument currently under development by the authors, constitute the measurement plan for the forthcoming pilot study. Directions for future work include: piloting the model at one or more Kazakhstani universities using a quasi-

experimental design; validating the scenario-based and portfolio instruments across institutional contexts; and conducting a comparative study with similar programs in Central Asian and post-Soviet countries.

VI. CONCLUSION

This paper has presented an AI-integrated three-component model for computer science teacher preparation, grounded at the intersection of international frameworks, sustainable development policy, and empirical data from a survey of 899 teachers across five regions of Kazakhstan. The model demonstrates that effective preparation of computer science educators in the age of AI cannot be limited to a purely technical focus. Technical, ethical, and social competencies must be developed simultaneously as interdependent dimensions of a unified professional system. The principle of structural simultaneity, embodied in the 3×4 matrix, ensures that all three dimensions are present at every stage of preparation.

The practical implications of the model include: (a) for curriculum developers – the 3×4 matrix as a tool for auditing and designing study plans; (b) for university administrators – a rationale for the necessity of interdisciplinary collaboration and institutional support; (c) for education policymakers – an argument for including ethical and social competencies in computer science teacher preparation standards.

The model can serve as a reference framework for education policy developers in Kazakhstan and in regions with similar conditions of digital transformation in education. Future work includes: experimental piloting of the model at a university; development of validated assessment instruments for all three dimensions, aligned with indicators of the DigCompEdu progression; and a longitudinal study of the effects of model implementation on teacher professional development and student learning outcomes.

REFERENCES

[1] OpenAI, “How people are using ChatGPT,” 2025. [Online]. Available: https://openai.com/index/how-people-are-using-chatgpt/
[2] UNESCO, AI and Education: Guidance for Policymakers. Paris: UNESCO, 2023.
[3] P. Mishra and M. J. Koehler, “Technological pedagogical content knowledge: A framework for teacher knowledge,” Teachers College Record, vol. 108, no. 6, pp. 1017–1054, 2006.
[4] C. Redecker and Y. Punie, European Framework for the Digital Competence of Educators: DigCompEdu. Luxembourg: Publications Office of the EU, 2017.
[5] OECD, OECD Digital Education Outlook 2023. Paris: OECD Publishing, 2023.
[6] Government of the Republic of Kazakhstan, “Education Development Concept for 2023–2029,” Resolution No. 249, Mar. 28, 2023.
[7] D. S. Touretzky, C. Gardner-McCune, F. Martin, and D. Seehorn, “Envisioning AI for K-12: What should every child know about AI?” in Proc. AAAI Conf. Artificial Intelligence, vol. 33, no. 01, 2019, pp. 9795–9799.
[8] UNESCO, Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO, 2021.
[9] D. A. Kolb, Experiential Learning: Experience as the Source of Learning and Development. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[10] D. A. Schön, The Reflective Practitioner: How Professionals Think in Action. New York: Basic Books, 1983.
[11] L. S. Shulman, “Knowledge and teaching: Foundations of the new reform,” Harvard Educational Review, vol. 57, no. 1, pp. 1–22, 1987.

[12] A. Yadav, C. Mayfield, N. Zhou, S. Hambrusch, and J. T. Korb, "Computational thinking in elementary and secondary teacher education," ACM Trans. Comput. Educ., vol. 14, no. 1, pp. 1–16, 2014.
[13] F. Nabi, W. Vortia, and E. Shardey, "Teacher readiness for AI and digital tools in K-12 classrooms," Int. J. Multidisciplinary Res., vol. 7, no. 5, 2025.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Gender Pay Gaps in Software Engineering: A Literature Review of Attrition, Measurement, and Structural Inequality

Marija Maček
Department of Theoretical and Applied Foundations of Information Sciences
University of Zagreb Faculty of Organization and Informatics
Varaždin, Croatia
mmacek@foi.hr
ORCID: 0009-0001-2177-4553

Abstract
The gender pay gap in software engineering is often discussed as a labor-market outcome. Yet, existing estimates are shaped by how gendered participation in the profession is defined and measured. This paper presents a literature review of empirical studies published between 2016 and 2026 on gender pay gaps in software engineering and closely related IT roles. Following PRISMA 2020 guidelines, 28 studies were selected from Scopus, Web of Science, and IEEE Xplore and analyzed across three dimensions: factors affecting women's retention in or exit from software engineering; reported causes and characteristics of pay disparities; and methods used to estimate these gaps. The review shows that conventional human capital variables do not fully account for the observed disparities, and that organizational and sociotechnical factors influence both wage outcomes and women's retention in the profession. Recurrent threats to validity include broad role definitions, mixed employment contexts, and limited treatment of sample selection and survival bias. By integrating evidence on pay, retention, and attrition, the paper reframes the estimation of the gender pay gap in software engineering as a sociotechnical and methodological problem.

Keywords: gender pay gap; software engineering; attrition; literature review

I. INTRODUCTION

Despite decades of progress and formal commitments to equality in the technology sector, software engineering continues to show persistent gender inequalities, especially in the form of a gender pay gap and low retention of women in the profession. These patterns carry social and economic consequences for individuals, organizations, and the wider software ecosystem. Empirical studies show that conventional human-capital variables, such as education and experience in the human-capital theory sense [1], do not fully explain pay disparities in software engineering [2], [3]. Instead, wage outcomes are also shaped by organizational and sociological mechanisms, including biased performance evaluations, limited career progression, gendered labor market segmentation, unequal care burdens, and sociotechnical exclusion [4], [5]. This broader perspective matters because software engineering is embedded in institutional and economic arrangements that shape access, progression, and exit. A central challenge is that most pay-gap studies rely on samples of those currently employed in the sector. When workforce participation is non-random, estimates may be distorted by sample selection and survival effects, potentially leading to underestimation of inequality [6], [7]. Research on attrition and return pathways in technology work supports this concern and shows that retention is substantively as well as statistically important [8]. Software engineering also deserves separate attention from broader ICT categories because its work organization, including project-based deadlines, continuous upskilling, and promotion systems tied to visibility and access to high-value projects, may affect women disproportionately [2], [4]. Evidence from broader ICT and labor-market research further suggests that these processes do not operate uniformly across all technology occupations [9], [10]. In addition, software engineering combines technical prestige with organizational informality. While meritocratic ideals may reduce

inequality in principle, informal evaluation and culturally coded notions of who "fits" the field can reproduce disadvantage [11], [12]. Findings on glass-ceiling dynamics and organization-specific inequalities likewise indicate that pay gaps should be interpreted within wider institutional structures rather than as isolated salary outcomes [13], [14]. From this perspective, ignoring attrition may make inequality less visible [10], [15], while research on comeback pathways and gendered career trajectories reinforces the same interpretation [8], [16].

Accordingly, this paper synthesizes contemporary research (2016–2026) on the gender pay gap in software engineering, with particular attention to how gendered participation, retention, and exits shape the production and interpretation of pay-gap evidence. Using the PRISMA 2020 protocol, the review integrates findings from 28 studies to answer three research questions: (RQ1) which factors influence women's retention or exit from software engineering, (RQ2) what are the reported causes and characteristics of the pay gap, and (RQ3) which methods are used to estimate it. By linking retention evidence with pay-gap estimation practices, the review highlights where current approaches may mismeasure inequality and outlines implications for both empirical software engineering and gender-and-technology research. It argues that the visible population of senior women in software engineering may be selective rather than representative, meaning that apparent improvements in outcomes may partly reflect cumulative filtering of the labor force rather than reduced inequality [8], [16]. The review contributes by offering a structured synthesis of recent evidence, integrating pay-gap and attrition research, and showing that methodological design choices shape the empirical visibility of inequality, particularly when studies define samples through current sector participation [17], [7]. In this paper, "gender pay gap" refers to differences in earnings between women and men, both unadjusted and adjusted for productivity-related factors; "sample selection bias" refers to bias arising when inclusion depends on participation in the sector; and "survival bias" refers to conclusions based only on those who remain in the profession. This interpretation is consistent with labor-market research on computer-science workers [2], methodological work on pay-gap decomposition [7], and the original logic of sample-selection correction models [6].

II. METHODS

This review followed PRISMA 2020 recommendations to structure identification, screening, eligibility assessment, and synthesis. The aim was to make the review process transparent and reproducible while retaining enough flexibility to interpret heterogeneous evidence on pay, attrition, and labor-market participation in software engineering. The paper asks three research questions: **RQ1)** Which factors influence women's retention in, or exit from, software engineering? RQ2) What are the reported causes and characteristics of the gender pay gap in software engineering? **RQ3)** Which empirical and analytical methods are used to estimate the gender pay gap in software engineering?

Searches were performed in Scopus, Web of Science (WoS), and IEEE Xplore databases. The search strategy combined software engineering role terms, gender terms, and pay/retention terms. Filters were applied to include studies published between 2016 and 2026, written in English, and published as journal articles or conference papers. Studies from unrelated domains and studies focusing on students or school populations were excluded to maintain a labor-market focus. [17]. For this review, the inclusion of retention, turnover, and labor-force participation terms is especially important because the review addresses gendered participation in technology work, not only wage estimation formulas. Inclusion criteria covered empirical studies addressing pay, wages, earnings, retention, turnover, or career progression in software engineering or closely related IT roles; studies reporting gender-disaggregated outcomes or gender pay gap estimates; and peer-reviewed journal or conference publications (2016-2026, English). Exclusion criteria covered student-only samples, unrelated disciplines, duplicates, and records with incomplete metadata that prevented reliable assessment.

Where studies combined software engineering with broad ICT categories, this was retained but explicitly coded as a source of interpretive heterogeneity in the synthesis.

Search results were de-duplicated and screened in two stages: (1) title/abstract screening, followed by (2) full-text eligibility assessment. In total, 711 records were initially identified (Scopus: 514; WoS: 186; IEEE Xplore: 11). After screening and eligibility checks, 28 studies were included in the final synthesis. Both authors participated in study selection and appraisal. Rather than calculating a formal inter-rater agreement coefficient, we used a qualitative consistency check on a shared subset of records and resolved uncertainties through discussion. This procedure strengthened interpretive consistency, although the absence of a reported agreement statistic remains a limitation. The study selection process is summarized in Figure 1.

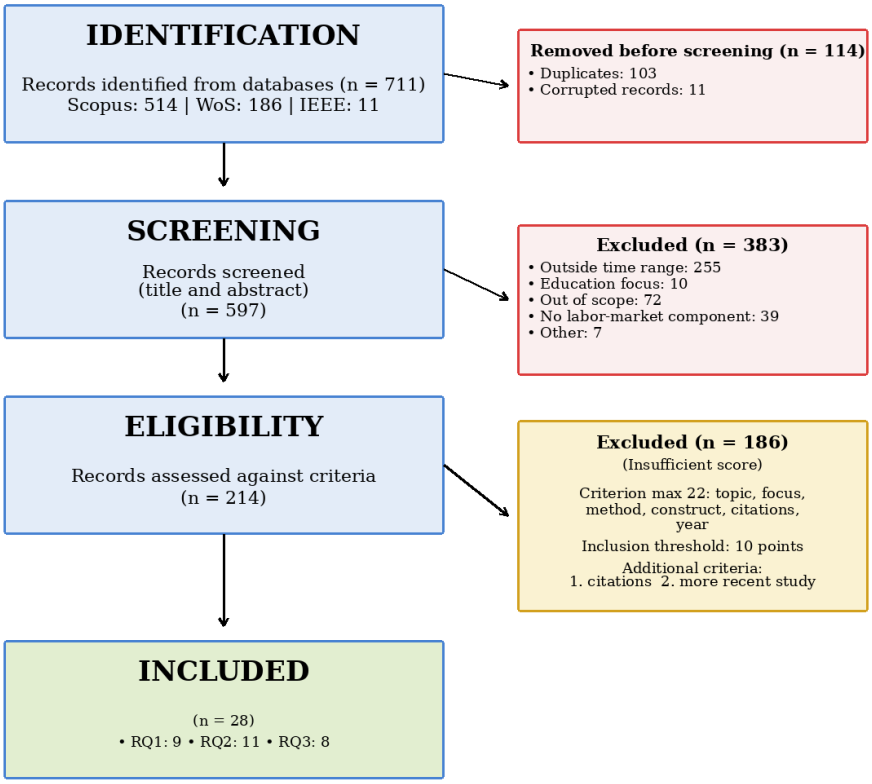


Figure 1. PRISMA 2020 literature selection flow diagram (Source: author's own elaboration)

To support the interpretation of heterogeneous evidence, a structured quality appraisal was performed at the study level. Given the diversity of study designs, the appraisal focused on role-scope definition (software engineering vs. broad IT), transparency of data and sample construction, control of key confounders, transparency of estimation methods, and discussion of attrition or non-participation. These criteria were selected because methodological work on wage gaps and studies of attrition in technology work both show that weak reporting can obscure how inequality is produced and observed [7], [6]. In this review, the appraisal also serves to identify how mechanisms of inequality are rendered visible or obscured by methodological choices. An evidence-weighting scheme (maximum 22 points) was retained as a secondary prioritization indicator (topic relevance, focus, method strength, theoretical construct, citations, and recency), while the quality appraisal functions as the primary basis for methodological interpretation.

For each included study, the review extracted country/region, population definition (SE vs. IT), employment context (corporate vs. freelance/self-employment where applicable), dataset type, sample size, outcome measure (pay/wage/earnings), estimation method (for example, OLS, Blinder-Oaxaca decomposition, quantile regression), controlled covariates, and key findings. Synthesis was narrative and thematic due to heterogeneity in populations, operationalizations, and methods, which precluded a meaningful quantitative meta-analysis. Themes were synthesized not only by method and outcome but also by the institutional and sociotechnical mechanisms that shape gendered participation and inequality. The review strategy was designed to balance software-engineering specificity with sufficient breadth to capture relevant adjacent IT roles discussed in the pay-gap literature. This choice reflects that many empirical studies do not clearly distinguish software engineers from software developers, ICT professionals, or broader technology workers. Rather than excluding such studies entirely, the review coded this domain ambiguity as an analytical issue in its own right [18], [9]. Broader comparative evidence on ICT work further shows why this distinction matters for interpretation [3]. A quantitative meta-analysis was not undertaken for both substantive and methodological reasons. The included studies differ substantially in the populations they examine, the operationalization of wages or earnings, the units of analysis, and the statistical models used to estimate gender differences. Some analyze corporate wage structures, others platform or freelance work, and others broader ICT labor markets, with only partial relevance to software engineering. Under such conditions, combining estimates into a single pooled effect would risk creating a misleading impression of comparability. Following PRISMA 2020 principles, the review instead emphasizes transparent documentation of selection and a narrative synthesis that foregrounds heterogeneity, methodological assumptions, and threats to validity [17]. The quality appraisal was also intended to go beyond a simple ranking exercise. In literature on gendered work, weak role definitions and poor reporting of sample construction can themselves be analytically significant because they shape which inequalities become visible in the evidence. Studies that do not separate freelancers from salaried employees, or that pool software engineers with broad ICT categories, may understate the organizational sources of inequality found in more role-specific analyses [19], [20]. The absence of discussion of attrition or non-participation adds a further interpretive limitation [7].

For language improvement, sentence-level restructuring, and grammatical checking, the author used ChatGPT (GPT-5.2 Thinking, OpenAI). The tool was used only to support wording, language refinement, and editorial consistency of author-written text. All substantive arguments, literature searching, study selection, appraisal, synthesis, interpretation, and final verification of the manuscript were performed and validated by the authors, in line with the publisher's AI policy [21].

III. RESULTS

A total of 711 records were identified across Scopus, Web of Science, and IEEE Xplore. After deduplication, screening, eligibility assessment, and evidence weighting, 28 studies were included in the final synthesis. The reviewed studies span multiple geographical contexts and combine journal articles and conference papers, with substantial variation in populations, data sources, and analytical methods. Cross-national evidence from Canada and Central Europe illustrates this heterogeneity [22], [18]. Additional studies from the Czech ICT context and broader IT settings point to similar variation in wage patterns and labor-market composition [23], [24]. Broader comparative work in IT also supports this reading [3]. This variation is analytically important because it reflects different labor regimes, occupational classifications, and institutional settings rather than only technical differences in datasets.

1. RQ1: Factors influencing women's retention in, or exit from, software engineering

The reviewed studies indicate that a non-linear interaction between structural constraints and organizational climate shapes women's retention in software engineering. Across contexts, retention is associated less with

pay levels alone and more with workplace inclusiveness, managerial support, promotion processes, and compatibility between job demands and family responsibilities [15], [10]. Several studies also highlight the protective role of role models and inclusive culture in reducing isolation and supporting persistence in the profession [12], [11]. These findings suggest that attrition is not random and should be understood as both a methodological concern and a gendered exclusion process within technology work. If women are more likely to exit due to adverse organizational or structural conditions, then the subgroup of women observed in wage datasets is positively selected [10], [15]. Return-to-work research in IT points to similar cumulative effects over time [8]. Across the reviewed studies, retention emerges as a cumulative process rather than a one-time decision. Women do not simply respond to one factor such as salary; rather, they navigate a layered environment in which work intensity, family demands, managerial support, promotion transparency, and role models jointly shape whether remaining in the field is perceived as sustainable [15], [8]. Evidence from firm-level studies further supports the importance of inclusive organizational cultures [12].

2. RQ2: Causes and characteristics of the gender pay gap in software engineering

The synthesis confirms that the gender pay gap is reported across multiple national and organizational contexts, although its magnitude and mechanisms vary. In many studies, a gap persists after controlling for conventional human capital variables, suggesting that differences in education and experience are insufficient to explain the observed disparities [22], [18]. The reviewed evidence instead points to structural mechanisms such as biased performance evaluation, glass-ceiling effects, and differential career progression [13], [5]. From a gender and technology perspective, these mechanisms should be read as part of broader gendered divisions of labor and organizational power. Professional norms and evaluation systems matter for advancement [13], [5], and they also interact with care-related constraints and occupational sorting in technology work [4], [2]. The results also show that the nature of the gap depends on employment context and job classification. Studies focusing on freelancing or platform-based work often report dynamics that differ from those observed in corporate settings [20], [19]. In organizations, compensation structures may also include opaque bonuses, promotion-related rewards, and evaluation-driven pay growth [25]. Wage gaps are also not static across contexts or career stages. In some national settings, gaps appear early among recent graduates or early-career professionals [22], [26]. In others, inequality becomes more pronounced through promotion systems, accumulation of seniority, or unequal access to high-reward roles [23], [27]. A further distinction concerns employment form. Studies focused on freelancing or firm ownership often find different gendered pay dynamics than studies of salaried corporate employment [20], [19]. These contrasts are analytically useful because they reveal which dimensions of inequality are specific to corporate hierarchies and which travel across labor-market arrangements [4].

3. RQ3: Methods used to estimate the gender pay gap

The dominant methodological approaches in the reviewed literature are regression-based models and decomposition techniques, especially the Blinder-Oaxaca decomposition. More recent studies extend this baseline through quantile regression, machine-learning-based prediction, and digitally oriented analyses of labor-market change [28], [27]. These methods improve descriptive granularity and may better capture heterogeneity across the wage distribution [29], [30]. Related work on digitalization and wage inequality also broadens the methodological context in which gendered pay outcomes are interpreted [31]. However, despite increasing methodological sophistication, explicit correction for sample selection remains rare in studies focused on software engineering and closely related IT roles. In practice, most models estimate pay gaps conditional on current participation in the sector, while treating labor-force presence as exogenous. This assumption is difficult to justify in a domain characterized by gender-differentiated retention and career interruptions [6], [7]. Attrition research in IT points to the same problem from a different angle [8]. From a gender and technology perspective, this methodological omission matters because it can depoliticize structural

attrition by treating exclusion from the field as analytically irrelevant to interpreting the pay gap. Methodologically, the literature demonstrates both progress and a persistent blind spot. On the one hand, the shift from simple mean comparisons to decomposition methods, quantile regression, and machine-learning approaches has improved the ability to distinguish between explained and unexplained components of wage disparities [28], [29]. On the other hand, many of these advances still operate on samples defined by current participation and therefore inherit the same underlying selection problem [30], [25]. This is where the absence of selection correction becomes especially consequential. Heckman's [6] formulation makes clear that when participation in the observed sample is non-random, estimated outcome equations may be systematically biased. Yet in the reviewed software-engineering and adjacent IT literature, the problem is more often acknowledged indirectly through concepts such as the leaky pipeline or career breaks than modeled directly in pay-gap estimation [8], [7].

4. Cross-study synthesis: recurrent threats to validity

Across RQ1-RQ3, five recurrent threats to validity emerge: (1) population-definition heterogeneity (software engineering vs. broad IT/ICT categories), (2) mixing of employment contexts (corporate employment vs. freelancing/self-employment), (3) incomplete treatment of compensation components, (4) limited modeling of attrition and non-participation, and (5) weak comparability of methods and covariate sets across studies. These issues recur across EU labor-market studies and work on platform or freelance arrangements [9], [20]. They also connect directly to methodological debates about correcting for composition and selection effects [19], [7]. These threats do not invalidate the literature; rather, they delimit what can be concluded and show how empirical visibility of inequality depends on methodological design choices.

IV. DISCUSSION

Table 1 synthesizes evidence across RQ1-RQ3 and makes explicit the methodological gap linking attrition, pay-gap estimation, and sample-selection/survival bias. The central contribution of this review is the identification of a methodological gap at the intersection of two research streams that are often studied separately: (1) women's retention and attrition in software engineering and (2) gender pay gap estimation. The former literature documents mechanisms that systematically affect who remains in the profession [15], [10], while the latter typically estimates pay differences among those currently employed [2], [27]. When these strands are not integrated, pay-gap estimates can be interpreted as if they were based on random samples of the original workforce. From a gender and technology perspective, this integration also reveals a sociotechnical inequality mechanism: institutions and work arrangements shape exits, and those exits, in turn, shape what later appears as evidence of equality.

Table 1. Research gap matrix linking findings from RQ1, RQ2, and RQ3 (Source: authors' own elaboration)

Factors influencing women's retention and exit in SE (RQ1)	**Causes of the gender pay gap (RQ2)**	**Applied methods (RQ3)**	**Identified research gap**
Organizational climate and support [15], [12]	Bias in performance evaluation [5]	Standard OLS regression [28], [32]	Sample selection bias OLS models estimated on a selected sample underestimate the true gap because they ignore women who left SE.

Psychological pressure [10], [11] Transparency [25]		Machine learning [31]	Lack of exclusion variables in models Psychological factors and transparency are rarely used as corrective factors in wage equations.
Family status [2], [4]	Glass ceiling [13], [33], [14]	Quantile regression [29], [30]	Survival bias Decomposition methods explain the gap only for the "survivors" in software engineering. The economic loss caused by exits from the sample is not estimated.
Career interruption [8]	Work arrangement [20], [19]	Comparative analysis of secondary data [9]	Ignoring the inactive population Standard wage analyses do not capture women temporarily outside the labor market.
Perception of the gender gap [9]	Global presence [3], [22], [18], [23], [24]	Blinder-Oaxaca [27] Resume/CV analysis [16]	Lack of correction Although theoretically recognized, a sample-selection correction method is not applied specifically to the software engineering domain.

The findings support stronger connections between empirical software engineering and gendered labor-process analysis. Future studies should report population scope with greater precision, distinguish software engineering roles from broader IT categories where possible, and explicitly discuss the implications of attrition and labor-force non-participation [9], [11]. The review also shows why structural barriers and wage dynamics should be interpreted together rather than separately [27]. The review also cautions against interpreting small observed pay gaps as evidence of equity when high differential attrition may already have reshaped the workforce and hidden cumulative disadvantage [10], [15]. Return pathways and re-entry studies suggest that this cumulative disadvantage is not always visible in cross-sectional pay estimates [8]. This review does not empirically test selection-correction models; however, it shows why such approaches are methodologically relevant for software engineering pay-gap estimation. In particular, the literature supports the need for future studies that jointly model workforce participation and wage outcomes, or otherwise assess the sensitivity of conclusions to plausible selection mechanisms [6], [7]. For gender and technology scholarship, this also implies a need for designs that better capture career interruptions, care-related constraints, and transitions between formal employment, freelance work, and non-participation [4], [2]. Comparative work on platform and self-employment settings suggests that these transitions may change how inequality is expressed rather than eliminate it [20], [19]. The review further highlights the value of richer longitudinal data and clearer reporting of role classifications, compensation structures, and career interruptions [27], [30]. It also supports closer attention to self-perception and invisible barriers in technology careers [16].

This review is limited by database coverage, publication-language restrictions, heterogeneity of included studies, and the absence of a quantitative meta-analysis. The synthesis also depends on the reporting quality of primary studies, which varies substantially across contexts and methods. In addition, because many studies conflate software engineering with broader ICT categories, some conclusions should be read as SE-centered rather than SE-exclusive. This is one of the central trade-offs of the review design: retaining broader ICT evidence improves coverage, but it also reduces precision when findings are interpreted as describing software engineering alone. The uneven maturity of the underlying literature also constrains the review's interpretation. Some studies use advanced quantitative methods but broad occupational categories, while others are more contextually specific but less directly comparable in statistical terms. In addition, the review necessarily relies

on what published studies report about their samples, variables, and model decisions; when attrition, bonus structures, or internal promotion criteria are not well described, this limits the precision with which one can assess the likely direction and magnitude of bias. These limitations do not undermine the value of the review, but they reinforce its central claim that stronger inference requires stronger reporting and more explicit treatment of participation dynamics. A further limitation is procedural. Both authors participated in study selection and appraisal, and uncertainties were resolved through qualitative discussion of a shared subset of records. However, the review did not calculate a formal inter-rater agreement statistic such as Cohen's kappa. This means that consistency was strengthened through collaborative judgment rather than quantified agreement, and the results should be read with that limitation in mind.

IV. CONCLUSION

This literature review shows that both wage-setting mechanisms and gendered retention and attrition shape the estimation of the gender pay gap in software engineering. By synthesizing evidence on pay disparities, exits from the profession, and analytical methods, the paper demonstrates that methodological choices can obscure or reveal inequality. Its main contribution is to show that pay-gap measurement, labor-force participation, and sociotechnical exclusion are analytically inseparable. The findings suggest that the issue is not only whether women are paid less than men, but also how software engineering is constituted as an observable labor market. Recent evidence from the broader IT sector also suggests that the interaction between experience and education may shape observed patterns in the gender pay gap [34].

REFERENCES

[1] Becker, G. S. (1964). Human capital: A theoretical and empirical analysis, with special reference to education. University of Chicago Press.

[2] Sassler, S., & Meyerhofer, P. (2023). Factors shaping the gender wage gap among college-educated computer science workers. PLoS ONE, 18(9), e0293300. https://doi.org/10.1371/journal.pone.0293300

[3] Saeed, K., Rubio-Herrero, J., & Mitra, A. (2024). Do women make less than men in IT? Yes and no! MIS Quarterly Executive, 23(1), Article 4.

[4] Tambe, P. B., & Yang, T. (2025). The hidden cost of IT innovation: Access to emerging technologies and the gender wage gap. MIS Quarterly, 49(1), 1-25. https://doi.org/10.25300/MISQ/2024/18268

[5] Zonno, R. (2026). Gender inequalities in performance evaluation: a systematic literature review. Equality, Diversity and Inclusion: An International Journal, 45(1), 1-22. https://doi.org/10.1108/EDI-04-2025-0289

[6] [6] Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47(1), 153-161. https://doi.org/10.2307/1912352

[7] Sprengholz, M., & Hamjediers, M. (2024). Intersections and Commonalities: Using Matching to Decompose Wage Gaps. European Sociological Review, 40(1), 1-17. https://doi.org/10.1177/07308884221141100

[8] Hyrynsalmi, S. M. (2024). Rebooting the System and Building New Futures: Supporting Women's Comeback in IT. In 2024 IEEE/ACM 5th Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering (GE@ICSE) (pp. 15-22). https://doi.org/10.1145/3643785.3648485

[9] Gawel, A., & Kapsdorferova, Z. (2024). Women in the ICT sector in European Union states: Facing gender inequalities. Studia Europejskie-Studies in European Affairs, 28(1), 89-107. https://doi.org/10.33067/SE.1.2024.6

[10] Luthra, A., Singh, K., & Anchal, M. (2023). Information overload as an antecedent of turnover intentions among software engineers of north Indian organizations: a study based on gender and marital

status. International Journal of Business and Globalization, 34(2), 194-212. https://doi.org/10.1504/IJBG.2023.134943

[11] Sánchez-Gordón, M., & Colomo-Palacios, R. (2024). On the Intersectionality of Software Practitioners and Role Models. In 2024 IEEE/ACM 5th Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering (GE@ICSE) (pp. 9-14). https://doi.org/10.1145/3643785.3648483

[12] Santos, A. L., Alexandre, J. D., & Meireles, F. (2021). A survey about gender diversity in a more gender-balanced IT firm. In 2021, the 4th International Conference of the Portuguese Society for Engineering Education (CISPEE) (pp. 1-7). https://doi.org/10.1109/CISPEE47794.2021.9507227

[13] Corbett, C., Wullert, K. E., Gilmartin, S. K., & Simard, C. (2024). Glass ceilings, step stools, and sticky floors: The racialized gendered promotion process. Socius: Sociological Research for a Dynamic World, 10, 1-18. https://doi.org/10.1177/23780231241274238

[14] Segovia-Pérez, M., Castro Núñez, R. B., Santero-Sánchez, R., & Laguna Sánchez, P. (2020). Being a woman in an ICT job: An analysis of the gender pay gap and discrimination in Spain. New Technology, Work and Employment, 35(1), 20-39. https://doi.org/10.1111/ntwe.12145

[15] Farooq, H., Janjua, U. I., Madni, T. M., Waheed, A., Zareei, M., & Alanazi, F. (2022). Identification and analysis of factors influencing turnover intention of Pakistan IT professionals: An empirical study. IEEE Access, 10, 31817-31829. https://doi.org/10.1109/ACCESS.2022.3181753

[16] Rapoport, E., Volkova, N. V., & Barajas, A. (2026). Women and men in tech: Invisible barriers, gendered self-perception, and the social reshaping of work. Technology in Society, 84, 103116. https://doi.org/10.1016/j.techsoc.2025.103116

[17] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

[18] Nedomová, L., Maryška, M., & Doucek, P. (2021). Gender Pay Gap in the Czech Information and Communication Technology Professionals. In Hradec Economic Days 2021 (pp. 572-582). https://doi.org/10.36689/uhk/hed/2021-01-059

[19] Kritikos, A. S., Maliranta, M., Nippala, V., & Nurmi, S. (2024). Does the gender of firm ownership matter? Female entrepreneurs and the gender pay gap. Journal of Population Economics, 37, 21. https://doi.org/10.1007/s00148-024-01030-x

[20] Liyanage, D., Vidanapathirana, A., Attanayake, P., Mudannayake, D., & Ambepitiya, K. (2025). Breaking barriers for females in IT freelancing: An analysis of challenges in Sri Lanka. Women's Studies International Forum, 109, 103173. https://doi.org/10.1016/j.wsif.2025.103173

[21] Taylor & Francis. (n.d.). AI policy. Retrieved February 10, 2026, from https://taylorandfrancis.com/our-policies/ai-policy/

[22] Jehn, A., Walters, D., & Howells, S. (2021). Employment and wage gaps among recent Canadian male and female postsecondary graduates. Higher Education Policy, 34(1), 184-206. https://doi.org/10.1057/s41307-019-00162-0

[23] Maryška, M., Nedomová, L., & Doucek, P. (2023). ICT Professionals' Wages Development - Is the Economy in a Resilience Period? In Proceedings of the 16th International Conference Liberec Economic Forum (pp. 366-375). https://doi.org/10.15240/tul/009/lef-2023-18

[24] Doucek, P., Maryška, M., & Nedomová, L. (2023). ICT Professionals in the Czech Business Before and During the Pandemic Era. In Proceedings of the 15th International Conference on Strategic Management and Its Support by Information Systems (SMSIS 2023) (pp. 40-48).

[25] Jaiswal, R., Gupta, S., & Tiwari, A. K. (2023). Dissecting the compensation conundrum: a machine learning-based prognostication of key determinants in a complex labor market. Management Decision, 61(9), 2685-2710. https://doi.org/10.1108/MD-07-2022-0976

[26] Nedomová, L., & Doucek, P. (2021). Impact of the first wave of Covid-19 on the trends of ICT Professionals' wages in the Czech Republic. Quality Innovation Prosperity, 25(1), 106-122. https://doi.org/10.12776/qip.v25i1.1468

[27] Prakash, A., & Yadav, I. S. (2025). Analyzing gender pay disparities and structural barriers among software developers: a cross-country study. Journal for Labor Market Research, 59(1), 1-18. https://doi.org/10.1186/s12651-025-00407-z

[28] Gupta, N., Singh, P., & Balcom, S. A. (2022). When pay equity policy is not enough: Persistence of the gender wage gap among health, education, and STEM professionals in Canada, 2006-2016. Canadian Studies in Population, 49(1), 1-24. https://doi.org/10.1007/s42650-022-00069-z

[29] Nguyen, D. T., Le, T. T. N., & Tran, V. Q. (2025). Exploring the impact of digital transformation and qualifications on wage disparities in urban and rural areas of Vietnam. Journal of Economics and Finance, 49, 1-22. https://doi.org/10.1007/s12197-025-09711-3

[30] Rahul, M., & Srivastava, D. (2022). The Gender Pay Gap Among Indian Software Developers. Journal of Economics, Race, and Policy, 5(4), 273-286. https://doi.org/10.1007/s41996-022-00099-4

[31] Lukyanova, A. L. (2021). Digitalization and the gender wage gap in Russia. Ekonomicheskaya Politika, 16(2), 88–117.

[32] Cruz, J., Bastos, A., Casaca, S. F., & Proença, I. (2025). Gender pay gap trends across STEM fields: an empirical analysis. International Journal of Sociology and Social Policy, 45(2), 1-17. https://doi.org/10.1108/IJSSP-02-2025-0090

[33] Prakash, A. (2025). What's in a name? Examining the impact of upper caste surname on wages and employment in the Indian information technology sector. Equality, Diversity and Inclusion: An International Journal, 44(1), 1-18. https://doi.org/10.1108/EDI-09-2024-0409

[34] M. Maček & M. Novak, “Gender Pay Gap Dynamics in the IT Sector: An Analysis of the Interaction between Experience and Education,” in INTED2026 Proceedings. Valencia: IATED, 2026, p. 0205. doi: 10.21125/inted. 2026.0205.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

AI-Driven Dynamic Trust Scoring over Self-Sovereign Identity Systems

Deniz Salucu
Industrial and Financial Technology
Information Communication Technologies
İstanbul, Turkey
deniz.salucu@turkcell.com.tr
ORCID: 0000-0003-0246-8777

Tunga Sayıcı
Digital Channel Solutions
Information Communication Technologies
İstanbul, Turkey
tunga.sayici@turkcell.com.tr
ORCID: 0009-0003-6150-8692

Abstract
Digital identity systems are under increasing pressure from sophisticated fraud mechanisms and rapidly evolving user behavior, exposing the limitations of traditional static verification models and centralized architectures. Although Self-Sovereign Identity (SSI) introduces a paradigm shift by enabling decentralized identity ownership through decentralized identifiers (DIDs) and verifiable credentials (VCs), it primarily addresses identity authenticity and lacks mechanisms for continuous trust evaluation. This limitation creates a critical gap in modern digital ecosystems, where verified identities can still exhibit risky or anomalous behavior. This paper proposes a novel AI-driven trust framework that augments SSI with dynamic trust scoring, transforming trust from a static verification outcome into a continuously adaptive metric. The proposed architecture integrates cryptographic identity validation with real-time behavioral analytics, incorporating identity confidence, behavioral consistency, transaction risk, and contextual signals into a unified and time-dependent trust function. Privacy is preserved through Zero-Knowledge Proof (ZKP) mechanisms, enabling secure, selective disclosure of identity attributes without exposing sensitive data. At the same time, blockchain technology ensures transparency and auditability via an immutable record of trust-related decisions. To achieve scalability and real-time responsiveness, the framework adopts an event-driven architecture supported by distributed and federated learning techniques, allowing trust models to evolve without compromising data sovereignty. The main contributions of this work are a formal, time-dependent, and multi-dimensional trust-scoring function that augments SSI; a privacy-preserving reference architecture that unifies event-driven analytics, federated learning, and blockchain-based auditability; and an evaluation methodology that defines the accuracy, latency, privacy, and scalability metrics against which the framework should be validated. By bridging the gap between identity verification and behavioral trust assessment, the proposed model introduces a generalized, interoperable, and privacy-preserving trust paradigm applicable to finance, healthcare, IoT, and digital governance, representing a significant step toward next-generation digital identity systems that are adaptive, context-aware, and inherently trustworthy.

Keywords: Self-Sovereign Identity (SSI), Decentralized Identifiers (DID), Verifiable Credentials (VC), Artificial Intelligence, Trust Scoring, Zero-Knowledge Proof (ZKP), Blockchain, Behavioral Analysis, Digital Identity, Privacy-Preserving Systems, Federated Learning

I. INTRODUCTION

The rapid digitalization of services in finance, healthcare, telecommunications, and public governance has significantly increased the importance of secure and reliable identity management systems. Digital identity is no longer merely an authentication mechanism; it has become a foundational layer enabling trust, access control, and secure interactions in distributed environments. However, traditional identity systems, which are predominantly built on centralized architectures and static authentication models, are increasingly inadequate to

address the complexity of modern digital ecosystems. These systems operate under the implicit assumption that once a user's identity is verified, trust can be continuously granted without reassessment. This assumption fails to account for the dynamic and evolving nature of user behavior, as well as the growing sophistication of fraud, account takeovers, and insider threats [1].

Self-Sovereign Identity (SSI) has emerged as a transformative approach that shifts identity ownership from centralized authorities to individuals. By leveraging decentralized identifiers (DIDs) and verifiable credentials (VCs), SSI enables users to control, manage, and selectively share their identity attributes without relying on intermediaries. This decentralized model enhances privacy, reduces dependency on centralized data stores, and minimizes the risk of large-scale data breaches. Furthermore, cryptographic primitives such as Zero-Knowledge Proofs (ZKPs) enable users to prove specific claims about their identity without disclosing the underlying data, aligning with modern data protection regulations such as GDPR [2] and reinforcing the principle of data minimization [3]. Despite these advantages, SSI systems are inherently limited in their ability to assess trust dynamically. They provide strong guarantees about identity authenticity, but do not evaluate behavioral trustworthiness over time. In real-world scenarios, a verified identity does not imply benign behavior; legitimate users can still engage in anomalous or malicious activities due to compromised accounts, insider threats, or evolving behavioral patterns. This limitation highlights a fundamental gap between identity verification and trust assessment. Artificial Intelligence (AI) offers powerful capabilities to bridge this gap by enabling real-time behavioral analysis, anomaly detection, and risk assessment. Machine-learning models can continuously learn from user interactions and identify deviations from expected behavior patterns. However, existing AI-driven trust systems are typically centralized, opaque, and difficult to audit, raising concerns about the transparency, fairness, and trustworthiness of the decision-making process [4]. To address these challenges, this paper proposes a unified framework that integrates Self-Sovereign Identity with AI-driven dynamic trust scoring and is supported by blockchain-based auditability. By redefining trust as a continuously evolving metric shaped by behavioral and contextual factors, the proposed architecture enables adaptive, privacy-preserving, and verifiable trust evaluation, introducing a new paradigm for digital identity systems in which trust is continuously assessed rather than statically assumed. The principal contributions of this paper are fourfold. First, we formalize trust as a time-dependent, multi-dimensional scoring function that augments SSI without weakening its privacy guarantees. Second, we propose a privacy-preserving reference architecture that unifies event-driven behavioral analytics, hybrid AI models, federated learning, and blockchain-based auditability within a single trust layer. Third, we describe a trust-portability mechanism that allows verifiable trust attributes to be reused across domains through Zero-Knowledge Proofs. Fourth, we define an evaluation methodology that specifies the accuracy, latency, privacy preservation, and scalability metrics by which such a framework should be validated, as well as its qualitative advantages over static and centralized alternatives.

II. METHODOLOGY AND TECHNICAL FRAMEWORK

1. Self-Sovereign Identity Foundations

Self-Sovereign Identity (SSI) represents a fundamental paradigm shift in digital identity management by redefining how identity is established, controlled, and verified in distributed environments. Unlike traditional identity systems that depend on centralized authorities for credential issuance, storage, and validation, SSI introduces a decentralized architecture in which individuals retain full ownership and control over their identity data. This transformation is enabled through the integration of decentralized identifiers (DIDs), verifiable credentials (VCs), and cryptographic verification mechanisms, effectively eliminating the need for trusted intermediaries and significantly reducing vulnerabilities associated with centralized data repositories, such as large-scale data breaches and single points of failure [1], [3]. Within this architecture, decentralized identifiers

function as globally unique, persistent, and cryptographically verifiable identity references. Each DID is associated with a DID Document that contains essential cryptographic materials, including public keys and service endpoints, enabling secure authentication without reliance on centralized identity providers. These DID Documents are typically anchored on distributed ledger technologies, allowing independent resolution and verification by relying parties in a trustless manner. Complementing this structure, verifiable credentials provide a standardized and interoperable framework for expressing identity attributes; they are issued by trusted entities, stored in user-controlled digital wallets, and can be presented without requiring real-time communication with the issuer, thereby ensuring both scalability and privacy preservation across heterogeneous systems [4].

A defining characteristic of SSI systems is their support for selective disclosure and Zero-Knowledge Proof (ZKP) mechanisms, which significantly enhance privacy and data minimization. Using ZKP protocols, users can prove the validity of specific claims without revealing sensitive information [5]. For example, instead of sharing a full identity record or disclosing personally identifiable attributes such as date of birth or income, users can cryptographically demonstrate compliance with a predefined condition, such as being over a legal age threshold. This capability aligns with modern regulatory requirements, including GDPR, and minimizes unnecessary exposure of personal data [2]. However, despite these strong guarantees of identity, authenticity, and privacy, SSI systems inherently lack mechanisms for dynamically evaluating trust. They effectively answer the question of identity, that is, who the user is, but do not address behavioral trustworthiness, that is, whether the user can be trusted at a given time. This limitation becomes increasingly critical in modern digital ecosystems characterized by identity theft, account compromise, insider threats, and behavioral fraud.

2. Dynamic Trust-Scoring Model

To address this gap, the proposed framework introduces an AI-driven trust-scoring layer that operates alongside SSI, transforming trust from a static attribute into a continuously evolving metric influenced by behavioral, transactional, and contextual factors. Trust is modeled as a time-dependent, multidimensional function that incorporates identity confidence, behavioral consistency, transaction risk, and contextual signals. Identity confidence reflects the strength and reliability of the underlying credentials, including the issuer's reputation, the credential's freshness, and the precision of biometric verification [6]. Behavioral consistency evaluates the alignment between current and historical user activity patterns, enabling the detection of anomalies indicative of fraudulent behavior. Transaction risk assesses the inherent sensitivity and potential impact of user actions, while contextual signals capture environmental factors such as device characteristics, geolocation, network conditions, and temporal patterns.

Formally, let the composite trust score of a user u at time t be defined over the set of trust dimensions D = {IC, BC, TR, CS, XD}, corresponding to identity confidence, behavioral consistency, transaction risk, contextual signals, and cross-domain trust, respectively. Each dimension is mapped to a normalized score φi(u,t) ∈ [0,1], and the composite trust score is expressed as:

$$T(u,t) = \Sigma_i \, w_i(c) \cdot \varphi_i(u,t) \,, \qquad \Sigma_i \, w_i(c) = 1 \qquad (1)$$

where wi(c) are context-dependent weights that sum to one and c denotes the interaction context. For the transaction-risk dimension, φTR is defined as one minus the normalized risk, so that higher values consistently denote greater trustworthiness. Because trust must reflect recent behavior more strongly than distant history, the behavioral-consistency dimension aggregates the per-event consistency of the user’s event history using an exponential time-decay weight:

$$\varphi_{BC}(u,t) = (\Sigma_k \alpha_k c(e_k)) / (\Sigma_k \alpha_k) , \quad \alpha_k = exp(-\lambda (t - t_k)) \qquad (2)$$

where tk is the timestamp of event ek, c(ek) is its instantaneous consistency, and $\lambda \geq 0$ controls the decay rate. The composite score is continuously updated as new events arrive, enabling real-time and adaptive decision-making. A policy engine compares T(u,t) against context-specific thresholds: when T(u,t) falls below a threshold τ, the system triggers additional safeguards such as step-up authentication or Zero-Knowledge-based re-validation, whereas high scores permit low-friction access. For example, a user who consistently accesses a system from a specific location and device may suddenly initiate access from an unfamiliar environment. However, the identity remains valid under SSI, the behavioral deviation lowers φBC and hence T(u,t), triggering re-validation. This represents a significant departure from binary authentication models, introducing a continuous trust-assessment paradigm in which every interaction contributes to an evolving assessment of user reliability. Table 1 summarizes the trust dimensions and their mapping to data sources and AI techniques.

Table 1. Trust-Scoring Dimensions and Technical Mapping

Dimension	Description	Data Source	AI Technique
Identity Confidence	Credential strength, issuer trust, and biometric validation	VC, DID, KYC systems	Rule-based + classification
Behavioral Consistency	Alignment with historical user behavior	User activity logs	LSTM, anomaly detection
Transaction Risk	Risk level of specific actions	Transaction systems	Supervised ML
Contextual Signals	Device, location, time, network patterns	Device telemetry, metadata	Feature engineering + ML
Cross-domain Trust	Trust portability across systems	External trust oracles	Graph-based learning

3. System Architecture

From a system architecture perspective, implementing this model requires a scalable, low-latency, and event-driven infrastructure capable of processing large volumes of streaming data. Figure 1 summarizes the proposed architecture and the flow of information among its components.

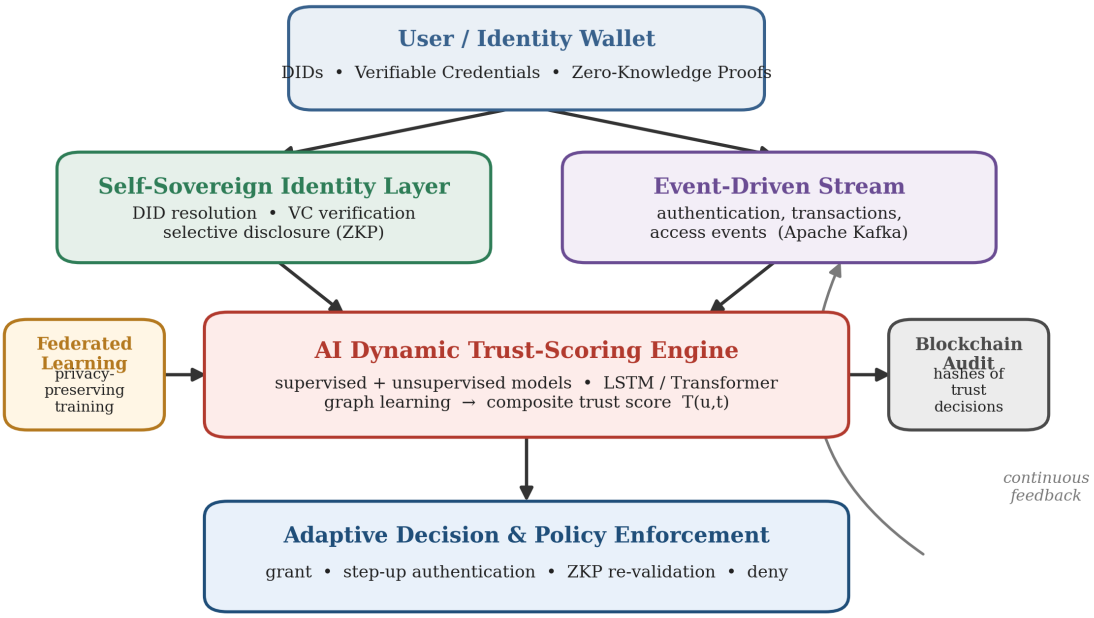


Figure 1. Reference architecture of the proposed AI-driven dynamic trust-scoring framework over Self-Sovereign Identity. *(system architecture)*

Each user interaction, such as an authentication attempt, a financial transaction, or a data access request, is treated as a discrete event and processed in real time by the trust evaluation engine. Events are ingested into a stream-processing pipeline, where machine-learning models analyze patterns, detect anomalies, and update trust scores with minimal latency. Technologies such as Apache Kafka and real-time processing frameworks enable efficient handling of high-throughput data streams, ensuring that trust assessments are performed without introducing system bottlenecks.

4. AI Modeling Layer

The AI layer in this architecture employs a hybrid approach that combines supervised and unsupervised learning techniques. Supervised models are trained on labeled datasets to identify known fraud patterns. In contrast, unsupervised models, such as anomaly-detection algorithms, can identify previously unseen threats by learning baseline behavioral patterns. In addition, sequence-based models such as Long Short-Term Memory (LSTM) networks [7] or Transformer architectures [8] can be employed to capture temporal dependencies in user behavior. Graph-based learning approaches further enhance the system by identifying complex relationships among entities, such as coordinated fraud networks or shared device usage patterns [9].

5. Privacy Preservation and Federated Learning

To address privacy concerns associated with centralized data processing, the proposed architecture incorporates federated learning, allowing machine-learning models to be trained across distributed data sources without transferring raw data to a central server [10]. This approach ensures data sovereignty and compliance with privacy regulations while maintaining model performance. Additional privacy-preserving techniques, such as

differential privacy [11] and secure multi-party computation, can further enhance the confidentiality of user data during model training and inference [12].

6. Blockchain-Based Auditability

Blockchain technology serves as a complementary component by providing an immutable and verifiable audit layer. Instead of storing sensitive identity or behavioral data on-chain, the system records cryptographic hashes of trust-related decisions, enabling independent verification without exposing the underlying information [13]. This design addresses the transparency limitations of AI-driven systems by providing a tamper-resistant record of decision-making processes, thereby increasing trust in both user interactions and system operations.

7. Trust Portability and Cross-Domain Interoperability

An additional dimension that extends the proposed framework beyond isolated identity and trust systems is the concept of trust portability and cross-domain interoperability, which fundamentally redefines how trust can be established, transferred, and reused across different platforms and organizational boundaries. In traditional architectures, trust is inherently siloed within individual systems, meaning that a user's reliability, reputation, or risk profile must be independently reassessed in each new environment. This leads to redundant onboarding processes, repeated verification steps, and significant inefficiencies in user experience. By leveraging the principles of Self-Sovereign Identity, trust can be partially decoupled from platform-specific contexts and instead associated with user-controlled credentials, enabling a more fluid and reusable trust layer [3].

Within this paradigm, trust is no longer confined to a single system but becomes a transferable attribute that can accompany the user across different domains. AI-generated trust scores or derived trust attributes can be embedded in verifiable credentials or dynamically generated through interoperable trust frameworks. For example, a user who has demonstrated consistent, low-risk behavior in a financial application may carry a cryptographically verifiable behavioral trust proof when interacting with a new service provider. Importantly, this proof does not expose raw behavioral or transactional data; instead, it relies on privacy-preserving constructs such as Zero-Knowledge Proofs or selective disclosure mechanisms to demonstrate that certain trust thresholds have been met [5].

This concept becomes particularly powerful in multi-domain ecosystems. Consider a scenario in which a user participates in a decentralized finance platform, a digital healthcare system, and an e-government service. Each of these systems may generate trust-related signals based on domain-specific interactions. Through an interoperable SSI framework, these signals can be abstracted into standardized trust indicators that can be reused across domains. For example, a high trust score in a financial system, reflecting consistent transaction behavior and low fraud risk, could contribute positively to access decisions in a separate system, such as granting expedited onboarding in a digital service platform. This reduces friction while maintaining a strong security posture. Similar cross-domain trust-propagation concepts have been explored in decentralized reputation systems and digital-trust frameworks [14], [15].

From a technical perspective, enabling trust portability requires defining standardized trust schemas and interoperable credential formats. Trust-related attributes must be encoded in a way that preserves semantic meaning across systems while remaining flexible enough to accommodate domain-specific variations. This introduces the need for trust-normalization mechanisms, in which scores generated in different environments are calibrated to a common scale or interpreted through context-aware weighting models. Without such

normalization, a trust score derived in one domain may be misinterpreted in another, leading to inaccurate or unfair decisions [14].

The role of decentralized reputation oracles further enhances this architecture by enabling the aggregation of trust signals from multiple independent sources. These oracles can act as intermediaries that collect anonymized behavioral indicators, compute aggregated trust metrics, and provide verifiable outputs to requesting systems. By operating in a decentralized manner, these oracles reduce reliance on single entities and mitigate the risk of manipulation or bias, while cryptographic guarantees preserve the integrity of aggregated trust data [13].

However, cross-domain trust sharing also presents several challenges. One primary concern is the risk of trust inflation or manipulation, in which users may attempt to artificially boost their trust scores in one domain to gain an advantage in another. To mitigate this, trust models must incorporate mechanisms for decay, contextual weighting, and anomaly detection. Additionally, the heterogeneity of trust requirements across domains necessitates adaptive scoring strategies that account for domain-specific risk profiles [15]. Privacy considerations are equally critical: techniques such as differential privacy, secure multi-party computation, and Zero-Knowledge Proofs can be employed to ensure that trust signals are shared without exposing underlying sensitive data [11], and user consent and control must remain central, allowing individuals to decide when and how their trust attributes are utilized across systems.

III. RESULTS AND DISCUSSION

The framework presented in this paper is architectural and conceptual; a full empirical evaluation on production-scale data is left for future work. Nevertheless, to make the proposal verifiable and comparable, this section defines the evaluation methodology, the metrics of interest, and the expected qualitative behavior of the system relative to existing approaches.

We envisage a prototype in which behavioral events are streamed via Apache Kafka, trust scores are computed by the hybrid AI models described in Section II, model updates are coordinated via federated learning across simulated organizational nodes, and decision hashes are anchored in a permissioned ledger such as Hyperledger Fabric. The framework should be evaluated along four dimensions. (i) *Detection quality* measures how accurately the trust score separates benign from malicious behavior, reported through precision, recall, F1-score, and the area under the ROC curve (AUC) on labeled fraud datasets and injected-anomaly scenarios. (ii) *Decision latency* measures the end-to-end time from event ingestion to trust-score update and policy decision, which must remain within interactive bounds, with a sub-second-per-event target, to support real-time authentication. (iii) *Privacy preservation* is assessed by verifying that raw behavioral data never leaves its origin node under federated training and by quantifying the utility–privacy trade-off introduced by differential-privacy noise. (iv) *Scalability* measures sustained event throughput and score-update latency as the number of users, events per second, and federated nodes increases.

Qualitatively, the proposed framework differs from prevailing approaches in three respects. Compared with static SSI verification, it adds a continuous, behavior-aware trust assessment, so that a valid credential no longer implies indefinite trust. Compared with centralized AI-based fraud-detection systems, it preserves data sovereignty through federated learning. It provides tamper-evident accountability through blockchain-anchored decision hashes, mitigating the opacity and single point of failure concerns of centralized models. Compared with decentralized reputation systems that expose reputation values directly, it discloses only Zero-Knowledge attestations that a trust threshold has been met, thereby preserving privacy during cross-domain reuse.

This analysis also surfaces the principal trade-offs. Federated learning and differential privacy improve privacy at some cost to model convergence speed and accuracy; blockchain anchoring adds auditability at the cost of additional write latency, which the design mitigates by recording only compact hashes rather than raw data; and cross-domain trust reuse improves user experience but requires trust-normalization and anti-inflation safeguards to remain sound. Establishing the quantitative operating points of these trade-offs, on representative datasets and workloads, together with an analysis of AI bias, explainability, and regulatory compliance, is the central objective of the planned experimental study.

IV. CONCLUSION

This study proposes a novel framework that integrates Self-Sovereign Identity (SSI) with AI-driven dynamic trust scoring to address the limitations of static and centralized identity systems. By distinguishing identity authenticity from behavioral trustworthiness, the model introduces a continuous trust-evaluation mechanism that adapts to real-time user behavior and contextual signals, formalized as a time-dependent, multi-dimensional trust function. The incorporation of Zero-Knowledge Proofs ensures strong privacy guarantees, while blockchain provides a transparent and verifiable audit layer, enabling accountability without exposing sensitive data. Together, these components establish a decentralized, privacy-preserving, and adaptive trust architecture suitable for modern digital ecosystems.

The proposed approach demonstrates that trust should be treated not as a fixed attribute but as a dynamic metric that evolves. While the framework offers significant advantages in security, flexibility, and interoperability across domains such as finance, healthcare, and IoT, challenges related to AI bias, explainability, and regulatory compliance remain. Guided by the evaluation methodology defined in Section III, future research will focus on prototype implementation and quantitative validation, improving model transparency, optimizing scalability, and developing interoperable trust standards. Overall, this work highlights a fundamental shift toward intelligent, context-aware trust models as a cornerstone of next-generation digital identity systems.

REFERENCES

[1] A. Tobin and D. Reed, “The Inevitable Rise of Self-Sovereign Identity,” The Sovrin Foundation, 2016.
[2] European Parliament and Council, “Regulation (EU) 2016/679 (General Data Protection Regulation),” Official Journal of the European Union, 2016.
[3] W3C, “Decentralized Identifiers (DIDs) v1.0,” World Wide Web Consortium, W3C Recommendation, Jul. 19, 2022.
[4] W3C, “Verifiable Credentials Data Model v2.0,” World Wide Web Consortium, W3C Recommendation, May 15, 2025.
[5] O. Goldreich, Foundations of Cryptography: Volume 2, Basic Applications. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[6] A. K. Jain, A. Ross, and S. Prabhakar, “An Introduction to Biometric Recognition,” IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 1, pp. 4–20, Jan. 2004.
[7] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[8] A. Vaswani et al., “Attention Is All You Need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[10] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.

[11] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Found. Trends Theor. Comput. Sci., vol. 9, nos. 3–4, pp. 211–407, 2014.

[12] K. A. Bonawitz et al., "Practical Secure Aggregation for Federated Learning on User-Held Data," in Proc. NIPS Workshop on Private Multi-Party Machine Learning, 2016.

[13] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008.

[14] D. Hardman, D. Reed, D. Chadwick, and M. Sporny, Introduction to Trust Over IP, Version 2.0. Trust Over IP Foundation, Nov. 17, 2021.

[15] M. Conti, E. De Cristofaro, and C. Soriente, "Trust in the Digital World," IEEE Commun. Surveys Tuts., vol. 20, no. 3, pp. 1846–1872, 2018.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Digital twin-based residual analysis for leak detection and localization in high-pressure gas pipelines

Murad Sharifov
Information Technology Department
Baku Higher Oil School
Baku, Azerbaijan
murad.sharifov.std@bhos.edu.az
ORCID: 0009-0009-8994-7316

Naila Allahverdiyeva
Information Technology Department
Baku Higher Oil School
Baku, Azerbaijan
naila.allakhverdiyeva@bhos.edu.az
ORCID: 0009-0002-0700-5472

Abstract
This work proposes developing a physics-based digital twin (DT) system for detecting and locating leaks in high-pressure gas transmission pipelines. The DT is based on the quasi-steady Weymouth flow equation and discretized across N = 20 spatial segments. First-order dynamics are introduced using time constants derived from the acoustic wave speed, thereby capturing dynamic behavior. Two complementary techniques for leak detection are used and compared: dual-channel fixed-threshold testing at the 3σ level and dual-channel cumulative sum (CUSUM) testing. To localize leaks, a normalized least-squares error surface is evaluated across all candidate pipeline segments, with sub-segment resolution obtained using parabolic interpolation. A 60-seed Monte Carlo simulation is performed on a 40 km pipeline with a nominal leak size of 8%. The CUSUM detector achieves an average detection delay of 50 s, compared with 80 s for the 3σ detector. The 3σ detector exhibits no false alarms, whereas the CUSUM detector has a false alarm rate of 3.3%. The average localization error is 1.77 km, corresponding to a single 2 km pipeline segment. Additional evaluation under a benchmark scenario consistent with the literature demonstrates that the proposed method produces realistic detection behavior across diverse pipeline conditions. This study provides a computationally efficient, interpretable, and low-cost solution for simultaneous leak detection and localization that requires only standard inlet and outlet pressure measurements.

Keywords: SCADA, Digital Twin, CUSUM, gas pipeline monitoring, leak detection, Weymouth equation, residual-based fault detection

I. INTRODUCTION

The safety and reliability of natural gas transmission lines are of utmost importance for ensuring energy security in both developed and developing countries today. The absence of gas leak detection systems can lead to direct economic losses from product loss, environmental damage from methane emissions, and the risk of catastrophic accidents from ignited gas. Multiple regulatory frameworks govern the operational requirements for leak detection systems, including the ENTSOG standards in Europe and the PHMSA integrity management program in the United States. Consequently, operators are required to have leak detection systems capable of determining the existence and location of any leak by comparing the amount of gas lost against “threshold” levels established by regulatory bodies. Recent digital twin approaches have demonstrated detection accuracy exceeding 90% for leak rates as low as 1–3%, yet these methods lack localization capability, which the present work addresses [1]. Conventional supervisory control and data acquisition (SCADA) systems generally use pressure monitoring heuristics that are inherently reactive: an alarm is triggered when a transient pressure event propagates through the pipeline to an instrumented node, causing the node to exceed a predefined threshold. A comprehensive review of leak detection methods based on digital twin (DT) confirms that current models primarily focus on detection rather than on identifying leak size or location [2]. These systems can have high

false-alarm rates and limited sensitivity to very small leaks (e.g., less than 1%- 2% of the nominal throughput). They cannot provide information on the spatial location of the fault unless the system is instrumented with densely spaced sensors. Transient model-based methodologies (hydraulic simulators that run in real time alongside the physical pipeline) do not share the same shortcomings as traditional SCADA systems. However, the principal challenge with transient model-based methodologies is ensuring that the hydraulic simulator accurately models the physical pipeline's behavior, so that model-reality residuals remain statistically stationary under fault-free conditions. The contributions of this work consist of three primary dimensions. First, we present our work, which offers a digital twin built from the quasi-steady Weymouth equation using physics-based first-order dynamics. This model was also verified against a noisy simulation of an actual system. Next, the second contribution to the literature includes the two separate detection methods used for event detection within the pipeline. These have been compared using a Monte Carlo simulation to determine the distribution of detection delay. Lastly, the third contribution evaluates a least-squares localization algorithm that uses parabolic interpolation across sub-segments. Collectively, these contributions provide pipeline operators with an interpretable, computationally efficient, and infrastructure-compatible framework for managing both detection and localization under realistic operating conditions.

II. BACKGROUND AND RELATED WORK

There is extensive information available on model-based detection of pipeline leaks. Begovich et al. illustrated that pressure-flow residuals from a steady-state hydraulic balance can be used to detect leaks in water systems. However, they lack temporal sensitivity to transients [3]. Ostapkowicz used cumulative sum tests to detect flow-imbalance signals in pipe systems, resulting in several-minute detection delays for liquid pipelines; however, no method for localizing leaks was proposed [4]. Liang et al. proposed a data-driven DT method using an autoencoder-CNN architecture trained on normal operating data from a real gas pipeline, achieving 97% accuracy but providing no leak localization [5]. Unlike data-driven approaches, the present work derives all model parameters from first principles, eliminating the need for historical leak samples. The current study addresses a very specific need by providing a physics-based detection method that is computationally feasible for real-time applications and provides both detection and localization using a unified, interpretable set of methodologies. In model-based systems, leak location is often treated as an inverse problem. Given a library of fault signatures, one for each possible proximity to the fault, the objective is to find the fault signature that best explains the observed residuals. In pipeline systems, each possible segment produces a unique pressure-flow profile at each observable node, allowing calculation of the least-squares error between the measured pressure-flow profile and each potential profile, thereby creating an error surface with a well-defined minimum. This defined minimum error value will correspond directly with the location of the leak [6]. Table I summarizes the key characteristics of representative prior works, along with the results of the present study.

Table 1. Comparison of Leak Detection Methods

Study	Method	Localization	Validation	Key Findings	Main Limitation
Al-Ammari et al. [2]	Hybrid DT + residual analysis	Yes	Real-field + experimental	100% detection, ARE < 3.21%	Mainly steady-state evaluation
Liang et al. [5]	AE-CNN digital twin	No	Real pipeline	R^2 = 0.9915, low FAR	No localization

Study	Method	Localization	Validation	Key Findings	Main Limitation
Liang et al. [7]	CNN-BiLSTM-AM residual model	No	Two real pipelines	Small-leak F2 = 88.04%	High computational complexity
Ebrahimi et al. [8]	Deep-learning DT framework	Yes	Simulation	>98% localization accuracy	Requires large simulated datasets; limited interpretability
This study	Physics-based DT + CUSUM	Yes	Simulation + Benchmark validation	50s average delay; ≈1 segment localization error	Experimental validation needed

III. MATHEMATICAL MODELLING

1. Quasi-Steady Weymouth Equation

Physics-based models remain foundational to pipeline integrity management, as data-driven approaches depend on sufficient historical anomaly data, which is rarely available in field operations [9]. Consider a pipeline of length L = 40 km and inner diameter D = 0.6 m, discretized into N = 20 equal segments of length $\Delta x = L/N = 2$ km. The Weymouth pressure-drop relation describes each segment:

$$p^2_{i+1} = p^2_i - K_{seg} \cdot q^2_i, \quad i = 0, 1, \ldots, N-1, \tag{1}$$

where p_i is the absolute pressure at node i, q_i is the local mass flow rate, and K_{seg} is the per-segment hydraulic resistance coefficient, which is defined as:

$$K_{seg} = f_D \cdot \Delta x \cdot Z \cdot R_g \cdot T_g / (D \cdot A^2), \tag{2}$$

with Darcy friction factor $f_D = 4 \times 0.0045$ (smooth-pipe approximation), compressibility factor Z = 0.90, specific gas constant $R_g = 518.3$ J/(kg·K), temperature $T_g = 293.15$ K, and cross-sectional area A. The pipeline is driven by a time-varying inlet pressure $p_{in}(t)$ and delivers flow to an atmospheric-referenced outlet. Nominal operating conditions correspond to an inlet pressure of 50 bar and a baseline mass flow of approximately 103 kg/s.

Because the Weymouth equation is algebraic, it yields quasi-steady-state (QSS) values $p_{qss}(t)$ and $q_{qss}(t)$ instantaneously for any given boundary conditions. Physical pipeline dynamics dominated by acoustic wave propagation are captured by first-order filters applied to both pressure and flow:

$$dp/dt = (p_{qss} - p) / \tau_p \tag{3}$$

$$dq/dt = (q_{qss} - q) / \tau_q \tag{4}$$

Critically, the time constants are not treated as free calibration parameters but are derived from the acoustic wave speed in the gas. For isothermal compressible flow, $c_s \approx 348$ m/s, giving:

$$\tau_p = L / c_s \approx 115\ s \tag{5}$$

$$\tau_q = \Delta x / c_s \approx 5.7\ s, \tag{6}$$

τ_p is the whole-pipeline wave travel time, and τ_q is the single-segment response time. This physics-based parameterization eliminates a degree of freedom that is problematic to identify from operational data alone and ensures that the dynamic model remains valid without re-tuning.

2. Leak Model

A leak at segment i_{leak} is represented as a point mass extraction: the downstream flow becomes

$$q_d = q_u - m_{leak}, \tag{7}$$

where q_u is the upstream mass flow, q_d is the downstream mass flow, and m_{leak} is the leaked mass flow rate [3]. In the main validation scenario, mleak = $0.08 \times q_b \approx 8.2$ kg/s (8% of baseline), placed at segment 12 (x = 24 km from inlet) and activated at $t_f = 1200$ s (20 min). The real system (RS) receives additive white Gaussian noise on both pressure and flow sensors with standard deviations $\sigma_p = 0.004 \times p_0$ and $\sigma_q = 0.004 \times q_b$.

IV. DIGITAL TWIN AND DETECTION FRAMEWORK

1. Parallel Twin Architecture

Two identical Weymouth models run in parallel at each sample instant, with the sampling period of $\Delta t = 2$ s. The measured inlet boundary conditions drive the digital twin (DT), but the DT receives no information about the leak. The real system (RS) comprises the same physical pipeline model, augmented by the active leak and measurement noise. The DT outputs the predicted pressure $\hat{p}(t)$ and mass flow $\hat{q}(t)$; the RS outputs the measured pressure pm(t) and mass flow $q_m(t)$. Residuals are formed at the pipeline outlet as follows [2]:

$$r_p(t) = p_m(t) - \hat{p}(t) \tag{8}$$

$$r_q(t) = q_m(t) - \hat{q}(t) \tag{9}$$

Under fault-free conditions, both residuals are zero-mean with standard deviations equal to the sensor noise levels. A leak shifts r_p negative (lower measured pressure than predicted) and r_q negative (lower measured flow than predicted) simultaneously, enabling a dual-channel detection strategy that is inherently more robust to single-channel noise exceedances than univariate tests.

2. Dual-channel 3σ Detection

The fixed-threshold detector compares both normalized residuals simultaneously against the 3σ level. A fault flag $F_{3\sigma}$ is asserted if and only if both conditions hold over at least two consecutive samples [10]:

$$|r_{p(t)}| > \varepsilon_p \ \ AND \ \ |r_{q(t)}| > \varepsilon_q \tag{10}$$

where thresholds are defined as:

$$\varepsilon_p = 3\sigma_p, \quad \varepsilon_q = 3\sigma_q \tag{11}$$

The dual-channel AND logic reduces the probability of a false alarm by squaring the single-channel false alarm probability:

$$P_{FA,dual} \approx P^2_{FA,single} \approx 9 \times 10^{-5} \tag{12}$$

This approximation assumes independence between channels and therefore represents a lower-bound estimate.

3. Leak Localization and Magnitude Estimation

Following detection, the localization is formulated as a model-based inverse problem. The digital twin is re-simulated for each candidate segment $i \in \{0, 1, \ldots, N-1\}$, with a hypothetical leak of estimated magnitude $\hat{m}_{leak} = \hat{q} - q_m$. This yields candidate predictions $p_{model}(t; i)$ and $q_{model}(t; i)$. The least-squares error at time t is defined as the sum over a sliding window W of normalized squared residuals (30 samples):

$$e(i) = \sum_{t \in W} \left[\frac{\left(p_m(t) - p_{\text{model}}(t,i)\right)^2}{\sigma_p^2} + \frac{\left(q_m(t) - q_{\text{model}}(t,i)\right)^2}{\sigma_q^2} \right] \quad \square\square\square\square$$

The normalized least-squares cost function (Eq. 13) is adapted from Al-Khomairi's inverse problem formulation and extended here to include dual-channel (pressure and flow) residuals [6]. The coarse segment estimate is obtained by minimizing the error surface:

$$i_{est} = arg\ min_i\ e(i) \tag{14}$$

$$x_{est} = (i_{est} + 0.5) \cdot \Delta x \tag{15}$$

Once the fault flag is asserted, the leak magnitude is estimated from the flow residual:

$$\hat{m}_{leak}(t) = \hat{q}(t) - q_m(t) \tag{16}$$

This estimate quantifies the missing mass flow between the predicted and measured signals. Additionally, this approach is consistent with the mass-conservation approach in [4]. Under quasi-steady conditions, it provides information on the accuracy of the approximation of the true leak rate. To reduce noise and improve stability, the estimate is smoothed with a short moving-average window.

V. EXPERIMENTAL RESULTS

1. Main Scenario

Figure 1 demonstrates the system response for an 8% leak at 24 km. Before the leak, the digital twin and the measured signals matched quite closely. Following the leak, however, both pressure and flow clearly diverged, resulting in negative residuals. About detection of the leak via the dual-threshold method, there is an elapsed time of approximately 80 seconds before a leak is detected. However, when using the cumulative sum (CUSUM) method, the elapsed time to detect the leak is reduced to 50 seconds. Thus, indicating that the leak is detected at a faster rate with CUSUM than with the dual-threshold approach.

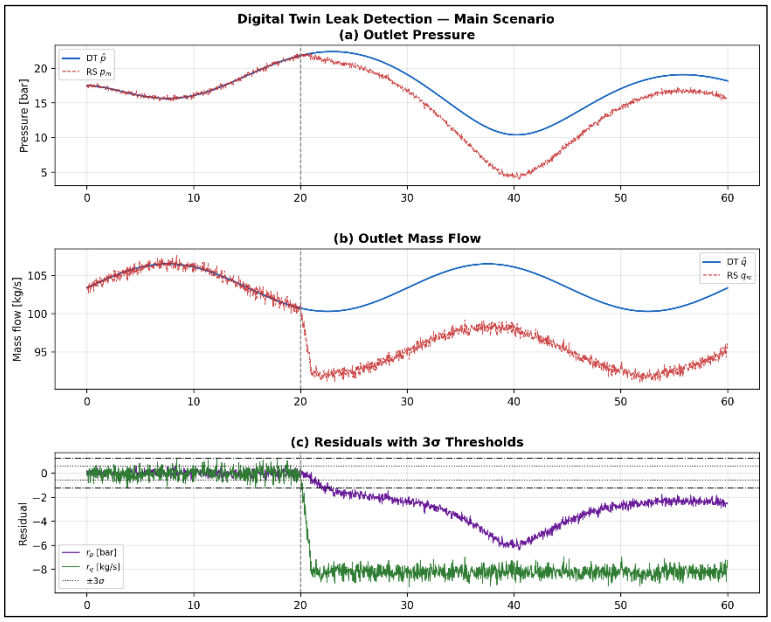


Figure 1. Main scenario. (a) Outlet pressure: DT prediction vs. noisy real-system measurement. (b) Outlet mass flow. (c) Dual-channel residuals with ±3σ thresholds

2. Monte Carlo Analysis

The accuracy of all detection and localization methods was assessed across 60 noise realizations to characterize the statistical distributions of relevant performance metrics. The average detection delay using the threshold-based method is 80 seconds, while the average detection delay using CUSUM is 50 seconds and exhibits statistically lower variability.

Figure 2(b) shows the distribution of the localization error obtained using the least-squares residual surface with sub-segment interpolation. In terms of leak localization accuracy, the localization error is predominantly within a single pipeline segment, centered at 1.77 km with a 95% confidence interval of [1.27, 2.16] km, thereby supporting consistent, accurate detection and localization regardless of noise.

Figure 2(c) summarizes the mean detection delays and associated 95% confidence intervals for both detection strategies. Overall, the Monte Carlo analysis indicates that the proposed digital twin framework delivers stable, statistically consistent leak detection and localization performance under measurement uncertainty.

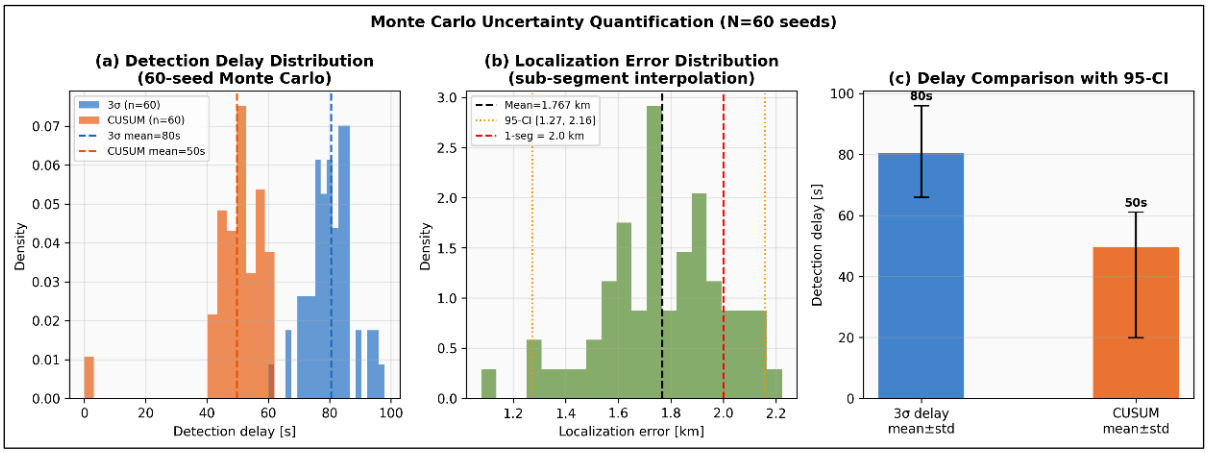


Figure 2. Monte Carlo uncertainty quantification. (a) Detection delay distributions for 3σ (mean 80 s) and CUSUM (mean 50 s). (b) Localization error distribution: mean 1.77 km, 95-CI [1.27, 2.16] km. (c) Delay comparison with ±1σ error bars.

3. Localization: LS Error Surface and Temporal Convergence

The localization algorithm converges to the correct region after the transient phase. While instantaneous estimates may fluctuate due to noise, temporal filtering stabilizes the result within 1 km of the true location. Figure 3 presents localization performance for the primary scenario examined. The LS reference error surface e(i) after t = 30 min is shown in panel (a). The location of the global minimum was segment 10; using parabolic interpolation gives an approximate new location estimate of 20.42 km. Given that the actual leak is located at segment 12 (x=24 km), with the addition of a median filter over the entire period after event onset, the final estimate is 23.258 km, corresponding to a localization error of 0.742 km. Figure 3(b) illustrates the time-series evolution of both unfiltered and filtered location estimates (represented by the orange and red colored dots).

To analyze generalization, the proposed method was qualitatively evaluated against published benchmark behavior, as per [4]. During this test, a 3% leak over 57 km was detected associated with this pipeline configuration. Under these circumstances, the CUSUM detected the leak in 134 seconds, which falls within the 3 to 8 minutes range reported by CERA [4]. On the contrary, the fixed 3σ threshold method took much longer, at 1736 seconds, to detect the leak, which reflects the low sensitivity of this method to leaks of that magnitude. Although the benchmark pipeline uses liquid flow, the acoustic-delay and flow-imbalance principles underlying the CUSUM detector are analogous, and the comparison is therefore indicative rather than directly equivalent. These results are consistent with existing literature indicating that cumulative detection methods are superior to fixed-threshold detection methods for detecting small-amplitude faults. Thus, it has been confirmed that the proposed framework yields realistic and comparable leak detection across various pipeline regimes.

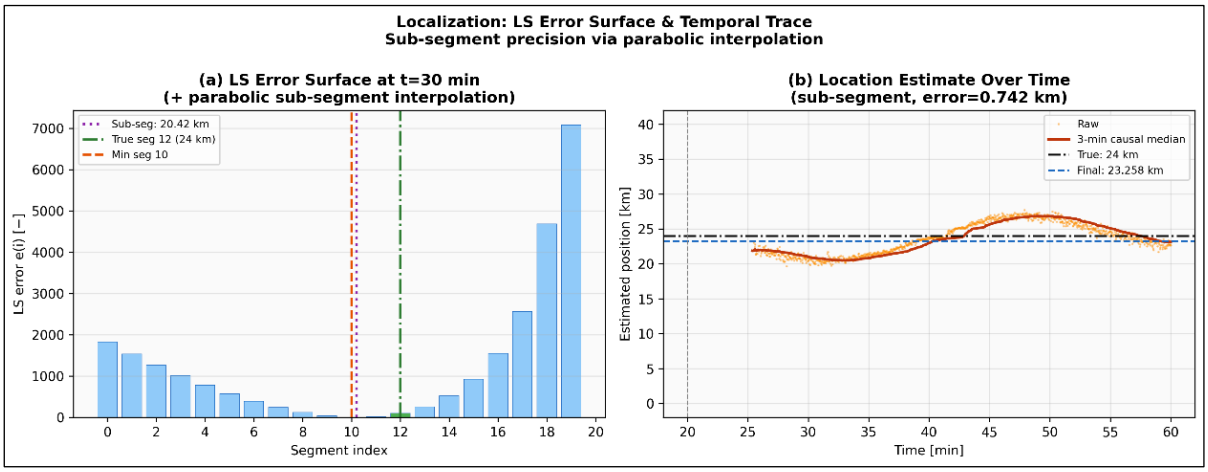


Figure 3. Localization performance. (a) LS error surface at t = 30 min, (b) Temporal trace of location estimate.

VI. DISCUSSIONS

All components of the proposed framework are derived from physical pipeline principles, providing transparency and interpretability for safety-critical applications. The CUSUM detector has reduced the detection delay from 80 seconds to 50 seconds, thereby demonstrating its utility for small leak detection. The accuracy of localization stays within the bounds of one segment (±2 kilometers) using only traditional inlet and outlet measurements, therefore making this technique practical to use with existing infrastructure. The performance of the localization is sensitive to model calibration; for example, a ±10% error in the friction coefficient would degrade accuracy, indicating the need for periodic calibration.

Several limitations of the present study warrant acknowledgment. The current work uses a simplified isothermal model, relies on simulated data, and assumes a single leak. The present framework assumes a single active leak. Numerical studies indicate that when two simultaneous leaks occur, the smaller leak becomes undetectable if its size is below 25% of the larger leak, a limitation that future extensions of this work should address [11]. Real-world variability, including changes in gas composition, was not included in the analysis. Additionally, the framework was validated exclusively on simulated data, which represents a significant limitation relative to field deployment. The CUSUM detector's performance is sensitive to the choice of drift parameter k and control limit h, which were empirically tuned for the simulated scenario; an automated tuning procedure would be required across different pipeline configurations. Furthermore, the localization accuracy reported here assumes a single active leak; the framework does not extend to multiple simultaneous leaks, a scenario that warrants future investigation. Future work will include parameter estimation, validation against actual data, and extension to multi-leak detection scenarios.

VII. CONCLUSION

A physics-based digital twin framework has been developed and validated for detecting and locating leaks in gas pipelines. This approach utilizes both residual analysis and two detection strategies, threshold and CUSUM, providing reliable detection with 50 to 80-second delays and relatively low false-alarm rates. Leak localization is performed using a least-squares scanning method, yielding an average error of 1.77 km along a 40 km

pipeline. The results indicate that the method works reliably under normal operating conditions; however, it is sensitive to model mismatch, underscoring the importance of regular calibration of the entire system.

In summary, the framework's overall results indicate that a physics-based approach to leak detection and localization offers a computationally efficient solution that does not require machine learning to produce accurate, interpretable results.

ACKNOWLEDGEMENT

Special thanks to Associate Professor Dr. Naila Allahverdiyeva for supervision and guidance throughout this research.

REFERENCES

[1] X. Cai and J. Wang, "A digital twin-driven natural gas pipeline leakage detection method," International Journal of Pressure Vessels and Piping, vol. 222, p. 105793, 2026.
[2] W. A. Al-Ammari, A. K. Sleiti, M. A. Rahman, S. Rezaei-Gomari, I. Hassan, and R. Hassan, "Digital twin for leak detection and fault diagnostics in gas pipelines: A systematic review, model development, and case study," Alexandria Engineering Journal, vol. 123, pp. 91-111, 2025.
[3] D. Aguiñaga, J. Alejandro, and O. Begovich, "Water Leak Diagnosis in Pressurized Pipelines: A Real Case Study," in Applied Condition Monitoring Modeling and Monitoring of Pipelines and Networks, 2017, pp. 235-262.
[4] P. Ostapkowicz, "Leak detection in liquid transmission pipelines using simplified pressure analysis techniques employing a minimum of standard and non-standard measuring devices," Engineering Structures, vol. 113, pp. 194-205, 2016.
[5] J. Liang, L. Ma, S. Liang, H. Zhang, Z. Zuo, and J. Dai, "Data-driven digital twin method for leak detection in natural gas pipelines," Computers and Electrical Engineering, vol. 110, p. 108833, 2023.
[6] A. Al-Khomairi, "Leak detection in long pipelines using the least squares method," Journal of Hydraulic Research - J HYDRAUL RES, vol. 46, pp. 392-401, 2008.
[7] J. Liang, S. Liang, L. Ma, H. Zhang, J. Dai, and H. Zhou, "Leak detection for natural gas gathering pipeline using spatio-temporal fusion of practical operation data," Engineering Applications of Artificial Intelligence, vol. 133, p. 108360, 2024.
[8] E. Ebrahimi, M. Kazemzadeh, and A. Ficarella, "Leak identification and quantification in gas network using operational data and deep learning framework," Sustainable Energy, Grids and Networks, vol. 39, p. 101496, 2024.
[9] M. Hussain, T. Zhang, R. Dwight, and I. Jamil, "Energy pipeline degradation condition assessment using predictive analytics – challenges, issues, and future directions," Journal of Pipeline Science and Engineering, vol. 4, p. 100178, 2024.
[10] M. Basseville and I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application. Prentice Hall, 1993.
[11] M. A. Adegboye, A. Karnik and W.-K. Fung, "Numerical study of pipeline leak detection for gas-liquid stratified flow," Journal of Natural Gas Science and Engineering, vol. 94, p. 104054, 2021.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Evaluating Quantum Computing Use Cases in Banking: A Sequential Model-Risk Admissibility Framework

Anandasubramanian Pranatharthy Codangudi
PhD Candidate, Department of Economics
VELS University
Chennai, India
anandcp@mac.com

Abstract
Quantum and hybrid methods are increasingly proposed for banking workloads, but theoretical speedups and favorable benchmarks do not establish that a regulated bank should pilot a given use case. What a bank needs is narrower: whether a specific candidate has enough surviving evidence to justify the next, more expensive stage of evaluation. This paper contributes a quantum-specific pre-pilot falsification framework built on four evidence objects: mechanism and advantage-source fit, classical benchmark sufficiency, encoding and end-to-end cost survivability, and governance admissibility. The gates are applied in a dependency-informed default sequence, distinguishing satisfied conditions, unresolved conditions, affirmative failures, and deployment-specific questions that cannot yet be assessed. Applied illustratively to the reviewed corpus, quantum amplitude estimation for pricing and risk is deferred; quantum optimization is deferred pending comparison against strong classical solvers; quantum machine learning that depends on quantum random access memory is rejected where dequantization applies; and near-term quantum machine learning is deferred. These are classifications based on the evidence reviewed, not on measured performance. The framework demonstrates no quantum advantage, has not received regulatory approval, and has not been empirically validated.

Keywords: quantum computing in finance, model risk management, classical benchmarking, pre-pilot screening, banking governance

I. INTRODUCTION

Banks use computationally intensive models for pricing, exposure assessment, risk management, optimization, and financial crime screening. Quantum and hybrid methods have been proposed for several of these workloads, but theoretical promise alone does not justify a bank pilot. Three claims must be separated. An algorithm may have an asymptotic advantage under stated assumptions; that advantage may or may not survive data preparation, encoding, execution, and a strong classical comparator; and an end-to-end advantage may still fail deployment requirements for validation, monitoring, controls, and reversibility. A bank, therefore, needs to know whether a specified candidate deserves the next unit of evaluation effort. The required evidence already exists in separate literatures: quantum mechanisms and resource estimates, classical benchmarking, encoding costs, model-risk governance, and staged investment under uncertainty. The gap is their integration into a pre-pilot decision protocol for regulated banking. This paper asks what evidence a quantum or hybrid banking use case must survive before further evaluation resources are committed. The proposed protocol uses four evidence objects: G1, mechanism and advantage-source fit; G2, strong classical benchmark sufficiency; G3, encoding and end-to-end cost survivability; and G4, governance admissibility. The default order is dependency-informed, not proven optimal. PASS permits normal advancement; CONDITIONAL permits diagnostic examination of later gates but remains binding; FAIL stops normal progression. Decisions are PROCEED, DEFER, REJECT, or REFRAME. The paper introduces no quantum algorithm or regulatory rule, proves no quantum advantage, and does not empirically validate the framework. SR 26-2 [1] and PRA SS1/23 [2] are governance reference points applied

within their stated scope and by analogy where appropriate. Sections II-III establish prior art and method, Section IV applies the framework, and Sections V-VII discuss implications, limitations, and conclusion.

II. RELATED WORK AND CONTRIBUTION GAP

1. Sequential Evaluation as Prior Art

Stage-Gate uses staged Go/Kill decisions [3]; real-options theory explains contingent commitment under uncertainty [4]; sequential experimentation values further information against acquisition cost [5]; and test-ordering theory shows that cost, rejection probability, and precedence can affect expected resource use [6]. These are prior art, not contributions of this paper. What they do not provide is quantum-banking test content: the correct classical comparator, whether data preparation destroys a speedup, or which governance requirements apply to a proposed banking model.

2. Pricing and Risk: Mechanism, Comparator, and Encoding

For pricing and risk, quantum amplitude estimation (QAE) offers $O(1/\varepsilon)$ expected-value query scaling, compared to $O(1/\varepsilon^2)$ for standard Monte Carlo under stated preparation assumptions [7], with applications to value-at-risk (VaR) and conditional value-at-risk (CVaR) [8]. For admissibility, however, the relevant comparison is stronger: quantum-accelerated multilevel Monte Carlo (MLMC) provides one such benchmark setting [9]. Resource estimates remain large: a representative derivative-pricing case requires about 8,000 logical qubits and T-depth 5.4×10^7 [10], reduced to about 4,700 qubits in a quantum-signal-processing formulation [11]. More importantly, Grover-Rudolph state preparation can eliminate QAE speedup for the distribution class studied [12], while mitigation is distribution-specific [13]. Thus, the query advantage does not establish an end-to-end advantage. Within this structured corpus, no peer-reviewed paper establishes that real bank pricing or risk distributions preserve the required state-preparation advantage end-to-end or convert resource estimates into bank-operational stop rules.

3. Optimization and Quantum Machine Learning

For optimization, comparator strength is binding. Quantum approximate optimization algorithm (QAOA) portfolio performance is sensitive to formulation, depth, and hardware [14]; quantum interior-point methods remain impractical at near-term scales [15]; and classical simulated annealing has outperformed quantum annealing in a studied multi-objective portfolio setting [16]. Hybrid annealing and constrained-QAOA studies provide evidence of implementation but do not establish an advantage over a modern mixed-integer programming (MIP) solver or an equally strong problem-specific baseline [17], [18]. For quantum machine learning (QML), input access and baseline choice are central. Quantum random access memory (QRAM)-dependent speedup claims carry strong practical assumptions [19], and dequantization shows that, for some such claims, equivalent classical data access removes the purported quantum advantage [20]. Classical data-rich models can also be competitive [21]; on finance data, preprocessing can drive apparent QML advantage [22], while quantum deep hedging at 16 qubits achieved comparable rather than superior accuracy [23]. No peer-reviewed paper identified in this corpus establishes a quantum advantage in optimization or QML over strong classical baselines on operationally realistic banking problems.



4. Hardware, Governance, and Evidence Boundaries

Hardware assumptions are dated inputs: noisy intermediate-scale quantum (NISQ) limitations and barren plateaus constrain variational methods [24], [25], while advances in fault tolerance can materially revise resource budgets [26].

SR 26-2, which superseded SR 11-7 and SR 21-8 in April 2026, supplies the current U.S. interagency model-risk framework within its stated scope [1], [27]; PRA SS1/23 supplies the UK reference point [2]. Neither creates quantum-specific validation rules. They are used here as governance principles, not as evidence of regulatory approval. The review is structured, not systematic: the corpus was assembled for framework construction, was current to May 2026, and used no preregistered search or independent second coding. Absence claims, therefore, mean only that no paper was identified in this corpus.

5. Surviving Contribution

The contribution is not staged screening, real options, sequential testing, mechanism analysis, benchmarking, resource estimation, or model-risk validation. It is their domain-specific integration into necessary-condition tests for a specified quantum banking candidate: mechanism source, strong classical comparator, end-to-end cost, and governance admissibility. The G1-G4 order is a dependency-informed default, not a proven cost-minimizing sequence. Test-ordering theory permits different orders when costs, rejection probabilities, and precedence differ [6]. This paper estimates none of those parameters; a cheap, decisive constraint may therefore be examined earlier.

III. METHODOLOGY: PRE-PILOT ADMISSIBILITY FRAMEWORK

1. Assessment Object and Sequence

The framework assesses a specified candidate, not a technology family. A candidate fixes seven elements: banking workload; claimed quantum mechanism and advantage source; data/input model; hardware assumption; classical comparator; operational time budget; and deployment boundary. A material change creates a new candidate. Thus, "QAE accelerates derivative pricing" is not assessable until the calculation, input/access model, hardware, comparator, accuracy/time requirement, and deployment boundary are specified. The default sequence is G1 - G2 - G3 - G4. PASS permits normal advancement. CONDITIONAL permits later diagnostic assessment but remains binding and blocks PROCEED. FAIL stops normal progression; later assessable gates are NOT REACHED. A cheap, decisive constraint may be examined earlier.

2. Four Evidence Gates

G1 - Mechanism and advantage-source fit. The candidate must identify the quantum mechanism, computational structure, hardware/input assumptions, and claimed source of advantage. FAIL requires no identifiable mechanism, a merely quantum label, or applicable evidence that the claimed advantage arises from asymmetric input access rather than the quantum mechanism. If the mechanism is genuine but a classical method matches it, the issue belongs to G2. G2 - Strong classical benchmark sufficiency. The claimed benefit must outperform a credible contemporary comparator for the same task, with matched data access, accuracy, number of instances, and operational constraints. Examples include MLMC with variance reduction for the QAE formulation [9], adjoint algorithmic differentiation (AAD) for suitable sensitivities [28], modern MIP- and problem-specific

solvers for optimization, and tuned classical classifiers. Missing decisive comparison is CONDITIONAL; matched evidence showing no benefit is FAIL.

G3 - Encoding and end-to-end cost survivability. The advantage must survive material preparation, loading, oracle/payoff construction, constraint encoding, connectivity, repeated execution, and, where applicable, measurement costs. Herbert shows why: state preparation can remove QAE speedup for the studied method and distribution class [12]. Unquantified material cost is CONDITIONAL; evidence that it negates the advantage is FAIL.

G4 - Governance admissibility separates class-level requirements from deployment evidence.

G4a asks whether applicable controls or structural constraints can be identified from supervisory/model-risk principles [1], [2]. It does not confer approval and may be examined early when a decisive constraint is known.

G4b asks whether a specific institution and implementation can provide required validation, monitoring, lineage, integration, fallback, third-party, and jurisdictional controls. Without an institutional deployment, it is NOT ASSESSABLE AT CLASS LEVEL, a scope designation rather than an adverse status.

Table 1. Operational Coding Rules

Gate	Test	Minimum evidence	Stop condition
G1 — Mechanism and advantage source	Is the claimed advantage attributable to a specified quantum mechanism under the stated input model?	Named mechanism; computational structure; hardware and input assumptions; claimed source of advantage	No identifiable mechanism; quantum label only; applicable evidence shows the claimed advantage comes from asymmetric input access rather than the quantum mechanism
G2 — Classical benchmark	Does the claimed benefit survive an appropriately strong classical comparator under matched conditions?	Credible contemporary comparator; matched data access, accuracy, instance, and operational constraint	**CONDITIONAL** if decisive matched comparison is missing; **FAIL** if matched evidence shows the claimed benefit does not survive
G3 — End-to-end cost	Does the claimed advantage survive the relevant computational pipeline?	Material preparation, loading, encoding, execution, and measurement costs were quantified where applicable	**CONDITIONAL** if a necessary material cost remains unquantified; **FAIL** if evidence shows the cost negates the claimed advantage

Gate	Test	Minimum evidence	Stop condition
G4a — Class governance	Can applicable controls or structural constraints be identified?	Relevant supervisory or model-risk principles and use-case characteristics	**CONDITIONAL** if material requirements remain unresolved; **FAIL** only where an identified structural constraint prevents the specified formulation
G4b — Deployment governance	Can the institution and implementation satisfy applicable requirements?	Deployment-specific validation, monitoring, lineage, integration, fallback, third-party, and jurisdictional evidence	**FAIL** only where required controls are demonstrably absent or inadequate; otherwise, outside scope when no deployment is specified

[a] Source: Author framework; gate criteria draw on [1], [2], [9], [12], and [28].

3. Coding and Decision Rules

Each assessable gate receives PASS, CONDITIONAL, FAIL, or NOT REACHED. PASS satisfies the necessary condition; CONDITIONAL marks unresolved evidence or assumptions; FAIL requires affirmative evidence that the condition does not hold; NOT REACHED follows an earlier FAIL. Missing evidence is therefore not converted into failure. Gate findings yield PROCEED, DEFER, REJECT, or REFRAME. PROCEED permits the next bounded step with no unresolved necessary condition; DEFER leaves a potentially satisfiable condition unresolved; REJECT records affirmative failure for the assessed formulation; REFRAME changes a material candidate element and therefore creates a new candidate.

4. Scope and Boundaries

The framework is an expert-review pre-pilot screen, not production validation, regulatory approval, procurement scoring, automated compliance, or a proven cost-minimizing algorithm. Whether it improves decisions, reduces evaluation expenditure, or yields reproducible classifications remains an empirical question.

IV. ILLUSTRATIVE APPLICATION

This section applies Section II evidence using Section III rules. Classifications are the author's analytical judgments of specified formulations, not measured performance, regulatory determinations, or empirical validation.

1. Quantum Amplitude Estimation for Pricing and Risk

For QAE pricing/risk, G1 is PASS because QAE has an identifiable mechanism and theoretical query-complexity advantage under stated assumptions [7], [8]. G2 is CONDITIONAL because the reviewed evidence does not establish a matched bank-operational comparison against the stronger MLMC setting [9]. G3 is CONDITIONAL because state preparation can eliminate the speedup under certain conditions [12], and mitigation is distribution-specific [13]; the corpus does not specify which condition characterizes the bank



workload. G4a is CONDITIONAL, and G4b is NOT ASSESSABLE AT CLASS LEVEL—decision: DEFER pending matched benchmarking and end-to-end encoding evidence.

2. Quantum Optimization

For portfolio optimization, G1 is CONDITIONAL because the broad formulation does not fix a single advantage source sufficiently for PASS [14]. G2 is CONDITIONAL: adverse classical comparisons exist [16], but no decisive matched comparison against an appropriate modern MIP or equivalent strong solver is established for the specified operational constraints. G3 and G4a are CONDITIONAL; G4b is NOT ASSESSABLE AT CLASS LEVEL. Decision: DEFER pending mechanism specification, strong benchmarking, and implementation-cost evidence.

3. Quantum Machine Learning

For the QRAM-dependent QML formulation to which the reviewed dequantization result applies, G1 is FAIL: comparable classical input access removes the claimed specifically quantum source of advantage [20]. This is not a claim against QML generally. G2-G4a are NOT REACHED; G4b is outside class-level scope. Decision: REJECT the formulation; a materially different input model or advantage source would be a new, reframed candidate. For broader near-term QML, G1 and G2 are CONDITIONAL. Classical models can be competitive [21], preprocessing can drive apparent advantage [22], and small quantum implementations can be comparable rather than superior [23]. Still, these findings do not affirmatively defeat every matched banking formulation. G3 and G4a are CONDITIONAL; G4b is NOT ASSESSABLE AT CLASS LEVEL. Decision: DEFER pending a specified mechanism, matched baseline, and end-to-end evidence.

4. Cross-Case Result

Across cases, the framework distinguishes affirmative failure, unresolved necessary evidence, and evidence outside the current unit of analysis. Only the first two affect the current decision; G4b non-assessability carries no adverse inference. Table 2 records the first failed or unresolved condition and the evidence required next; this is a framework implication, not evidence that the framework improves decisions.

Table 2. Illustrative Pre-Pilot Classifications

Use case	G1	G2	G3	G4a	G4b*	Decision	First binding issue
QAE pricing/risk	PASS	CONDITIONAL	CONDITIONAL	CONDITIONAL	Not assessable	DEFER	Strong matched comparator
Quantum optimization	CONDITIONAL	CONDITIONAL	CONDITIONAL	CONDITIONAL	Not assessable	DEFER	Mechanism-to-benefit specification
QRAM-dependent QML†	FAIL	NOT REACHED	NOT REACHED	NOT REACHED	Not assessable	REJECT / REFRAME	Advantage-source assumption
Near-term QML	CONDITIONAL	CONDITIONAL	CONDITIONAL	CONDITIONAL	Not assessable	DEFER	Mechanism and strong baseline

[b.] Source: Author classifications based on the evidence reviewed in Section II and cited in Section IV.

V. DISCUSSION

1. The Decision Object Is Evidence Survival, Not Algorithmic Advantage

Algorithmic, end-to-end, and bank-deployable advantage are different claims. Evidence does not automatically transfer between them. QAE illustrates the distinction: its mechanism survives G1 while the strong comparison and state-preparation questions remain unresolved. Optimization and near-term QML are deferred for incomplete comparative or implementation evidence; only the specified QRAM-dependent QML formulation has an affirmative G1 failure. The framework, therefore, separates FAIL, CONDITIONAL, and deployment-level non-assessability. A single readiness score can collapse these different evidentiary states and obscure what must be tested next.

2. Sequential Falsification and the Limits of the Ordering Claim

Sequential assessment does not prove that earlier gates are always cheaper. Its narrower rule is that unresolved or failed necessary conditions should not be treated as satisfied. A proof of concept cannot establish comparative advantage without a matched baseline, and circuit performance cannot establish end-to-end advantage when the material-loading cost remains unresolved. Stage-Gate, real options, sequential experimentation, and test-ordering already provide the general logic [3-6]. This paper contributes to the quantum-banking content of the tests, not new decision theory. G1-G4 is only a dependency-informed default; a formal ordering model could rationally reorder gates for particular candidates.

3. Governance Is a Distinct Evidence Object, Not a Final Veto

G4 is not a final bureaucratic veto. G4a identifies class-level controls or structural constraints and may be relevant at intake; G4b concerns institution-specific validation, monitoring, lineage, integration, fallback, and third-party controls that published literature cannot establish. The sequence is analytical, not an organizational workflow.

4. Practical Implication

Operationally, the framework first forces a candidate to specify its seven elements, then asks which necessary condition has failed, remains unresolved, or lies outside the current scope. That answer determines the next evidentiary task. This is a framework implication, not evidence that the protocol improves decisions or reduces cost.

VI. LIMITATIONS AND RESEARCH AGENDA

The framework has four principal limitations. First, the review is structured rather than systematic, was current to May 2026, and used no preregistered search or independent second coding; absence claims apply only to this corpus. The same corpus constructs and illustrates the framework, so Section IV is not an independent validation.

Second, classifications are single-author judgments. No inter-rater study establishes reproducibility, especially for candidate definition, comparator strength, material end-to-end costs, or the boundary between unresolved evidence and affirmative failure. Third, the default ordering is not optimized: test costs, rejection probabilities, information value, and precedence could justify a different sequence [5], [6]. Fourth, G4b cannot be established without a specific institutional implementation, and changes to hardware, algorithms, classical baselines, or deployment assumptions may require reassessment. Validation should therefore proceed in four directions: independent reviewers coding identical evidence packets; retrospective application to documented quantum-finance pilots without using eventual outcomes during initial coding; prospective comparison with existing bank evaluation processes; and a formal value-of-information/test-ordering model. Relevant outcomes include agreement, evidence requested, decision reversals, evaluation effort, and decision time. These are proposed tests, not demonstrated benefits. They may validate, modify, reorder, or reject parts of the present protocol.

VII. CONCLUSION

Quantum banking candidates should not be advanced based on theoretical speedups or favorable benchmarks alone. A specified formulation must establish a defensible source of quantum advantage, withstand an appropriate strong classical comparator, preserve that benefit through material end-to-end costs, and identify the governance requirements relevant to deployment. The paper contributes a quantum-specific pre-pilot falsification protocol integrating those evidence objects as necessary-condition tests. The illustrative cases distinguish affirmative failure from unresolved evidence and deployment-level non-assessability: QAE, optimization, and near-term QML are deferred in the reviewed formulations, while the narrowly specified QRAM-dependent QML formulation is rejected where the applicable dequantization result undermines its claimed source of advantage. The classifications are analytical, the gate order has not been proven optimal, and the framework has not been empirically validated. Its bounded claim is therefore simple: a plausible quantum mechanism is a reason to investigate, not a reason to pilot; escalation requires the specified banking use case to survive the relevant evidence tests.

REFERENCES

[1] Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency, "Revised guidance on model risk management [Attachment to SR 26-2]," April 17, 2026. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/SR2602a1.pdf

[2] Prudential Regulation Authority, "Model risk management principles for banks (Supervisory Statement SS1/23)," Bank of England, 2023. [Online]. Available: https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss

[3] R. G. Cooper, "The Stage-Gate idea-to-launch system," in Wiley International Encyclopedia of Marketing. Chichester, U.K.: Wiley, 2010. doi: 10.1002/9781444316568.wiem05014

[4] L. Trigeorgis and J. J. Reuer, "Real options theory in strategic management," Strategic Management. J., vol. 38, no. 1, pp. 42–63, 2017. doi: 10.1002/smj.2593

[5] J. J. J. Thijssen and D. Bregantini, "Costly sequential experimentation and project valuation with an application to health technology assessment," J. Econ. Dyn. Control, vol. 77, pp. 202–229, 2017. doi: 10.1016/j.jedc.2017.01.016

[6] D. Berend, R. I. Brafman, S. Cohen, S. E. Shimony, and S. Zucker, "Optimal ordering of independent tests with precedence constraints," Discrete Appl. Math., vol. 162, pp. 115–127, 2014. doi: 10.1016/j.dam.2013.07.014

[7] P. Rebentrost, B. Gupt, and T. R. Bromley, "Quantum computational finance: Monte Carlo pricing of financial derivatives," Phys. Rev. A, vol. 98, no. 2, art. 022321, 2018. doi: 10.1103/PhysRevA.98.022321

[8] S. Woerner and D. J. Egger, "Quantum risk analysis," npj Quantum Inf., vol. 5, art. 15, 2019. doi: 10.1038/s41534-019-0130-6

[9] D. An, N. Linden, J.-P. Liu, A. Montanaro, C. Shao, and J. Wang, "Quantum-accelerated multilevel Monte Carlo methods for stochastic differential equations in mathematical finance," Quantum, vol. 5, art. 481, 2021. doi: 10.22331/q-2021-06-24-481

[10] S. Chakrabarti, R. Krishnakumar, G. Mazzola, N. Stamatopoulos, S. Woerner, and W. J. Zeng, "A threshold for quantum advantage in derivative pricing," Quantum, vol. 5, art. 463, 2021. doi: 10.22331/q-2021-06-01-463

[11] N. Stamatopoulos and W. J. Zeng, "Derivative pricing using quantum signal processing," Quantum, vol. 8, art. 1322, 2024. doi: 10.22331/q-2024-04-30-1322

[12] S. Herbert, "No quantum speedup with Grover-Rudolph state preparation for quantum Monte Carlo integration," Phys. Rev. E, vol. 103, no. 6, art. 063302, 2021. doi: 10.1103/PhysRevE.103.063302

[13] A. C. Vazquez and S. Woerner, "Efficient state preparation for quantum amplitude estimation," Phys. Rev. Applied, vol. 15, no. 3, art. 034027, 2021. doi: 10.1103/PhysRevApplied. 15.034027

[14] S. Brandhofer, D. Braun, V. Dehn, G. Hellstern, M. Hüls, Y. Ji, I. Polian, A. S. Bhatia, and T. Wellens, "Benchmarking the performance of portfolio optimization with QAOA," Quantum Inf. Process, vol. 22, no. 1, art. 25, 2023. doi: 10.1007/s11128-022-03766-5

[15] A. M. Dalzell, B. D. Clader, G. Salton, M. Berta, C. Y. Lin, D. A. Bader, N. Stamatopoulos, M. J. A. Schuetz, F. G. S. L. Brandão, H. G. Katzgraber, and W. J. Zeng, "End-to-end resource analysis for quantum interior-point methods and portfolio optimization," PRX Quantum, vol. 4, no. 4, art. 040325, 2023. doi: 10.1103/PRXQuantum. 4.040325

[16] E. Aguilera, J. de Jong, F. Phillipson, S. Taamallah, and M. Vos, "Multi-objective portfolio optimization using a quantum annealer," Mathematics, vol. 12, no. 9, art. 1291, 2024. doi: 10.3390/math12091291

[17] S. N. Morapakula, S. Deshpande, R. Yata, R. Ubale, U. Wad, and K. Ikeda, "End-to-end portfolio optimization with hybrid quantum annealing," Adv. Quantum Technol., vol. 9, no. 4, art. e00753, 2026. doi: 10.1002/qute.202500753

[18] P. Thomassin, G. Guérard, S. Djebali, and V. M. Lambert, "A quantum model for constrained Markowitz modern portfolio using slack variables to process mixed-binary optimization under QAOA," Quantum Mach. Intell., vol. 7, no. 2, art. 102, 2025. doi: 10.1007/s42484-025-00330-z

[19] S. Aaronson, "Read the fine print," Nature Phys., vol. 11, no. 4, pp. 291–293, 2015. doi: 10.1038/nphys3272

[20] E. Tang, "Dequantizing algorithms to understand quantum advantage in machine learning," Nature Rev. Phys., vol. 4, no. 11, pp. 692–693, 2022. doi: 10.1038/s42254-022-00511-w

[21] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, "Power of data in quantum machine learning," Nat. Commun., vol. 12, art. 2631, 2021. doi: 10.1038/s41467-021-22539-9

[22] J. Mancilla and C. Pere, "A preprocessing perspective for quantum machine learning classification advantage in finance using NISQ algorithms," Entropy, vol. 24, no. 11, art. 1656, 2022. doi: 10.3390/e24111656

[23] E. A. Cherrat, S. Raj, I. Kerenidis, A. Shekhar, B. Wood, J. Dee, S. Chakrabarti, R. Chen, D. Herman, S. Hu, P. Minssen, R. Shaydulin, Y. Sun, R. Yalovetzky, and M. Pistoia, "Quantum deep hedging," Quantum, vol. 7, art. 1191, 2023. doi: 10.22331/q-2023-11-29-1191

[24] J. Preskill, "Quantum computing in the NISQ era and beyond," Quantum, vol. 2, art. 79, 2018. doi: 10.22331/q-2018-08-06-79

[25] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, "Variational quantum algorithms," Nature Rev. Phys., vol. 3, no. 9, pp. 625–644, 2021. doi: 10.1038/s42254-021-00348-9

[26] S. Bravyi, A. W. Cross, J. M. Gambetta, D. Maslov, P. Rall, and T. J. Yoder, "High-threshold and low-overhead fault-tolerant quantum memory," Nature, vol. 627, no. 8005, pp. 778–782, 2024. doi: 10.1038/s41586-024-07107-7

[27] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, "Guidance on model risk management (SR 11-7)," 2011. [Removed from Federal Reserve Online].

[28] L. Capriotti and M. Giles, "15 years of Adjoint Algorithmic Differentiation (AAD) in finance," Quant. Finance, vol. 24, no. 9, pp. 1353–1379, 2024. doi: 10.1080/14697688.2024.2325158

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A New Topological Index-Based Solution Approach for the Traveling Purchaser Problem

Aysun Yurttaş Güneş
Department of Mathematics
Bursa Uludag University
Bursa, Türkiye
ayurttas@uludag.edu.tr
ORCID:0000-0000-0000-0000

Hacer Özden Ayna
Department of Mathematics
Bursa Uludag University
Bursa, Türkiye
hozden@uludag.edu.tr
ORCID:0000-0003-1556-3511

İlker Kücükoğlu
Department of Industrial Engineering
Bursa Uludag University
Bursa, Türkiye
ikucukoglu@uludag.edu.tr
ORCID:0000-0002-5075-0876

Ismail Naci Cangül
Department of Mathematics
Bursa Uludag University
Bursa, Türkiye
cangul@uludag.edu.tr
ORCID:0000-0002-0700-5774

Abstract
In this study, we propose a graph-theory-based solution approach to one of the well-known routing problems, the traveling purchaser problem (TPP), using topological graph indices. The main aim is to measure the relationship between suppliers and products in TPP by calculating new topological indices to be developed here based on graph-theoretical methods. These novel topological indices, called the affordability index and the closeness index, generate efficient routes and purchasing plans for purchasing different types of products from suppliers located in different locations. The solutions obtained using topological graph indices are compared with those obtained without graph theory in the literature, and the advantages of this graph-theoretic approach over existing methods are revealed through these benchmarks.

Keywords: traveling purchaser problem, topological indices, affordability index, closeness index

I. INTRODUCTION

In this study, the traveling purchaser problem (TPP), which is a variation of the traveling salesperson problem (TSP) frequently studied in the literature, is considered. One of the most important sub-branches of graph theory in the last century is the study of topological graph indices. In this field, a daily event can be modeled with a suitable graph, and this model can be studied either manually or using computer programming languages, with mathematical arguments and tools. A graph consists of vertices and edges. In real-life problems, we can model people, institutions, streets, intersections, metro or bus stops, companies, cities, countries, cells, neurons, factories, landmasses, etc., with vertices, and the relationships between them with edges. For example, we can draw lines between two people who know each other and determine the popularity of individuals in a group or network by looking at these lines. When we model these kinds of relationships with graphs, we transform our everyday problems into mathematical problems that we can solve using mathematical methods. Using matrices, determinants, vectors, numbers, polynomials, functions, etc., corresponding to these graphs, we first determine some numbers for the graphs. Then, by interpreting them, we can access valuable information that would normally require a lot of money, human resources, time, and effort to obtain in an everyday problem. There are several tools and methods to study a graph mathematically. Graph parameters and graph matrices are two of them. Nevertheless, a more frequent and popular way to study graphs is through various topological graph indices defined in recent years. The first examples are Wiener [1] and Gutman and Trinajstic [2]. Nowadays,

there are over 4,000 such indices defined for various purposes, most of which are mathematical or chemical. These indices, which are nothing more than mathematical formulas, can be used to classify graphs, derive their properties, make approximations, or comment on their corresponding applications [3-5]. Studying the effects of adding or removing vertices and edges on various graph properties yields valuable insights. For example, when calculating the effect of removing a vertex or edge, repeatedly applying this operation yields smaller graphs, usually with pre-existing properties. By working backward, the properties of the large graph under investigation can be obtained by applying the formula we derived from the properties of the small graph obtained at the end. This method was used to calculate the effects of adding and removing edges on Zagreb indices in [6]. A topological graph index is a formula that is generally defined by means of vertex degrees, distances, or other graph parameters, and often contains simple mathematical expressions. To apply this formula, a data set from the graph is needed. This data set is nothing more than a numerical table that can be given depending on the graph's vertex partition, edge partition, distance partition, or another graph parameter. TPP is known to be NP-hard in terms of solution difficulty because it encompasses decisions related to the traveling salesperson problem, facility location selection, and cluster covering problems [7]. Therefore, many solution approaches have been developed in the literature for solving TPP. These solution approaches are divided into two groups: exact solution methods that guarantee an optimal result and heuristic methods that aim to find a near-optimal result. Among the methods that give exact solutions, branch-and-bound, branch-and-cut, and branch-and-price algorithms are frequently used [8]. In addition to these approaches, dynamic programming was used by [9], and constraint programming by [10]. These methods, which guarantee optimal results, have been considered solution methods for many TPP variants. On the other hand, many heuristic methods have been developed to solve TPP and produce optimal or near-optimal solutions with shorter processing times. In this study, a heuristic approach based on graph theory is developed to solve the TPP. It aims to measure the relationship between suppliers and products in TPP using new topological index calculations to be developed based on Graph theory. The heuristic approach is used for supplier selection in TPP, and supplier routes are generated based on the index values. The solutions obtained using the topological graph index or indices are compared with those obtained without graph theory in the literature, and the advantages of this graph-theoretic approach over existing methods are revealed. The ability to model with graphs has broadened their application areas.

II. PROBLEM DEFINITION AND MODEL FORMULATION

The TPP is a well-known extension of the TSP, in which a purchaser satisfies the demand for a set of products from a set of markets. At each market, several products are available at a given price and capacity. To satisfy the demand for products, the purchaser starts their tour, visits several markets, and returns to the depot. Unlike the TSP, the purchaser does not have to visit all markets if they can satisfy the demand for the products from a subset of the markets. Additionally, the purchaser can meet the demand for a product across multiple markets, since the number of available products is limited in each market. Within the restrictions above, the purchaser aims to minimize total procurement and transportation costs. In this context, the model formulation of the TPP is as follows.

Assumptions

- The tour of the purchaser starts and ends at the depot node.
- The purchaser visits several markets to satisfy the demand for several products, where each market can be visited once at most.
- The available quantity of a product at a market can be zero or less than the product demand due to the capacity restriction of the market.

Parameters

K Set of products
M Set of markets
M_k Set of markets in which the product k is available ($M_k \subseteq M$), $k \in K$
$\{0\}$ Depot node
V Set of markets and depot node ($M \cup \{0\}$)
d_k Demand amount of product k, $k \in K$
q_{ik} Available amount of product k at market i, $k \in K$, $i \in M_k$
p_{ik} Price of product k at market i, $k \in K$, $i \in M_k$
c_{ij} Traveling cost from node i to node j, $i, j \in V$
γ Large number

Decision variables

x_{ij} Binary variable and equals 1 if the purchaser travels from node i to node j, otherwise 0; $i, j \in V$, $i \neq j$
Yy_i Binary variable and equals 1 if mthe purchaser, otherise 0 visit market i.therise 0; $i \in M$
z_{ik} Amount of product k purchased from market i; $k \in K$, $i \in M_k$
u_i Integer variable used in the MTZ sub-tour elimination; $i \in$ V

Model

$$Min \sum_{i \in V} \sum_{j \in V} c_{ij} x_{ij} + \sum_{k \in K} \sum_{i \in M_k} p_{ik} z_{ik} \quad (1)$$

s.t.

$$\sum_{i \in M_k} z_{ik} = d_k \qquad k \in K \quad (2)$$

$$z_{ik} \leq q_{ik} y_i \qquad k \in K, \qquad i \in M_k \quad (3)$$

$$\sum_{\substack{j \in V \\ i \neq j}} x_{ij} = y_i \qquad i \in M \quad (4)$$

$$\sum_{\substack{i \in V \\ i \neq j}} x_{ij} = y_j \qquad j \in M \quad (5)$$

$$\sum_{i \in M} x_{i0} = 1 \quad (6)$$

$$\sum_{i \in M} x_{0i} = 1 \quad (7)$$

$$u_j - u_i \geq 1 - \gamma(1 - x_{ij}) \qquad i \in M, \qquad j \in V, \qquad i \neq j \quad (8)$$

$$x_{ij} \in \{0,1\} \qquad i, j \in V, \qquad i \neq j \quad (9)$$

$$y_i \in \{0,1\} \qquad i \in M \quad (10)$$

$$z_{ik} \geq 0 \qquad k \in K, \qquad i \in M_k \quad (11)$$

$$u_i \geq 0 \qquad i \in V \tag{12}$$

Equation (1) describes the model's objective function, which aims to minimize the purchaser's total procurement and transportation costs. Constraints (2) ensure that the demand of each product type must be satisfied. Constraints (3) provide that the quantity of a product purchased from a market cannot exceed the quantity of that product available in the market. These constraints also ensure that if a product is purchased from a market, that supermarket must be within the purchaser's route. Constraints (4) and (5) guarantee that a market can be visited once by the purchaser if a procurement is made at the corresponding market. Constraints (6) and (7) assure that the tour of the purchaser starts and ends at the depot node. Constraints (8) define the MTZ sub-tour elimination constraints. Finally, constraints (9)-(12) identify the model's decision variables.

III. METHODOLOGY

The proposed graph indices-based solution approach consists of two stages. In the first stage of the approach, the subset of the markets to be visited in the solution is formed by using two new topological indices. Following the first stage, the purchaser's route is determined across the selected markets. In this stage, the simulated annealing (SA) algorithm, one of the effective metaheuristic algorithms for routing problems, is employed. The following subsections provide details of the proposed solution approach.

1. Market Selection

In TPP, the purchaser can satisfy product demand through a subset of markets. In this context, the selection of markets to be visited by the purchaser is a critical decision layer for the problem, as it directly affects procurement and transportation costs. To make an efficient selection, two different topological indices are introduced to identify the offer-cost-based and transportation-cost-based relations between the markets. The indices are called the affordability index and the closeness index, respectively. The affordability index measures how affordable a market is relative to other markets in the problem. To calculate a market's affordability index, the prices of products available in that market are compared with those in other markets. For each product k at a market i, first the competitor market (i) cis found, where the competitor market is the market that sells the corresponding product at the lowest price i. c, k son alt simge, this means that market i. is selling product k at a better price, and this provides a benefit to market i. proportional to the number of products available in the market. Conversely if $p_{ik} > p_{i^c k}$, there is another market that sells it at a better price than market i. In this context, the affordability index of a market is determined by using Equation (13).

$$Aff_i = \sum_{k \in K} \begin{cases} (p_{i^c k} - p_{ik}) \times min\{q_{ik}, d_k\}, & if \quad p_{ik} < p_{i^c k} \\ (p_{i^c k} - p_{ik}) \times min\{q_{i^c k}, d_k\}, & if \quad p_{ik} > p_{i^c k} \end{cases} \quad i \in M \tag{13}$$

As the second topological index for the TPP, the closeness index measures the average distance of a market from a given number of other markets. For a market, i, the closeness index is determined by using the following equation,

$$Clos_i = \sum_{j \in M^c} c_{ij} \qquad i \in M \tag{14}$$

where M^c is the subset of M and consists of the markets closest to the market i..he size of the M^c is set to $\sqrt{|M|}$.

Once the markets are evaluated using an index, they are ranked according to the index value, which is denoted by büyük ha M üssü büyük hrf R. The markets re ranked in descending order f Aff_i and ascendig order of $Clos_i$. Following the ranking procedure, the market selection is made by including the markets from M^R into the solution until all product demand is satisfied. The set of selected markets is described as M^*.

2. Route Planning

Once the market selection is completed in the first stage, the TPP becomes the TSP, as procurement decisions have been eliminated. To determine an effective route plan for the purchaser, the SA is used to find the optimal order of visits to the selected markets. In the proposed SA, the initial solution is constructed using the nearest-neighborhood search algorithm, one of the best-known and most effective route-construction heuristics. After the initial solution is generated, the SA iteratively searches the solution space by using local search procedures. In each iteration of the SA, a new solution ($\boldsymbol{X'}$) is generated from the current solution ($\boldsymbol{X}$). In the proposed SA, the order of the selected markets is changed by using four different moves: exchanging a pair of markets (see in Figure 1a), moving a market to a forward position (see Figure1b), moving a market to a backward position (see in Figure 1c), and reversing a sub-path in the route (see Figure 1d). In each iteration, one of the local search procedures is randomly selected to apply to the current solution. $\boldsymbol{X'}$ is accepted according to the SA solution acceptance criterion where $\boldsymbol{X} = \boldsymbol{X'}$ if $f(\boldsymbol{X'}) < f(\boldsymbol{X})$ or $e^{-\Delta/T} \geq rnd$. Here, Δ is the total cost difference between the new solution and the existing solution ($\Delta = f(\boldsymbol{X'}) - f(\boldsymbol{X})$), T is the temperature, and rnd is the standard uniform distributed random number.

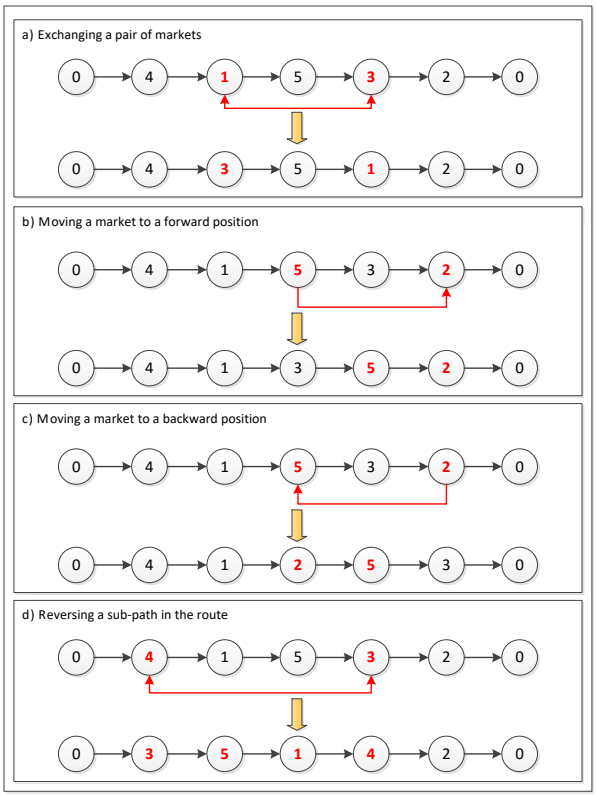


Figure 1. Local search procedures used in the proposed SA.

At the end of the main loop of the SA, the temperature is decreased by using a cooling coefficient (c). Additionally, the best-found solution ($\boldsymbol{X^{Best}}$) is updated if the newly generated solution is better than the current best-found solution. The SA terminates if the algorithm reaches the maximum number of iterations. The pseudo code of the proposed SA is given in Figure 2.

Input $\boldsymbol{M^*}$, c_{ij}
Set the algorithm parameters $T = T_0, c$
Generate an initial solution $\boldsymbol{X}$, $\boldsymbol{X^{Best}} = \boldsymbol{X}$
Do //Main loop of the algorithm
    Randomly select a local search procedure and generate $\boldsymbol{X'}$
    $\Delta = f(\boldsymbol{X'}) - f(\boldsymbol{X})$
    If $\Delta < 0$ **Or** $e^{-\Delta/T} \geq rnd$ **Then**
        $\boldsymbol{X} = \boldsymbol{X'}$
        If $f(\boldsymbol{X}) < f(\boldsymbol{X^{Best}})$ **Then**
            $\boldsymbol{X^{Best}} = \boldsymbol{X}$
        End If
    End If
    $T = T \times c$
Loop Until (Stopping criterion is met)
Output $\boldsymbol{X^{Best}}$

Figure. 1. Pseudo code of the proposed SA.

IV. COMPUTATIONAL RESULTS

To evaluate the performance of the proposed solution approach, a problem set of 180 instances is randomly generated. The generated problems are characterized by four metrics: the number of nodes, the absolute value of V, the absolute value of the end, and the absolute value of K. Absolute values of the end and of two levels (i.e., 10 and 50) are considered. The instances consisting of 10 nodes are called small-sized instances, while the instances with 50 nodes are called large-sized instances. The coordinates of each node are randomly generated between 0 and 100, and the distance between each pair of nodes is determined as the Euclidean distance. For the $\bar{C}_L$ and $\bar{D}_L$, three levels (i.e., 1, 2, and 3) are considered, where the price of each product in each market and its demand are randomly generated for the corresponding levels. For each combination of $|V|$, $|K|$, $\bar{C}_L$, and $\bar{D}_L$, five different instances are generated. Each instance is solved using the proposed approach, which accounts for both the affordability and closeness indices. Additionally, each instance is solved using the GUROBI solver with a 1-hour time limit to obtain a benchmark value. At the end of the time limit, the best-found integer solution is used. The results obtained by the proposed approach are compared with the GUROBI solution using Equation 15, where fvalue denotes the result obtained by the solution approach. The results obtained are presented in Tables 1 and 2 for the small- and large-sized instances, respectively. Each row in the tables presents the average percentage gap of five instances for the corresponding problem configuration.

$$Gap_{Index}\% = \frac{f_{Index} - f_{GUROBI}}{f_{GUROBI}} \times 100\% \quad (15)$$

Table 1. Obtained Results for the Small-Sized Instances

$\|K\|$	$\bar{C}_L$	$\bar{D}_L$	$Gap_{Aff}\%$	$Gap_{Clos}\%$
10	1	1	2.84	3.60
		2	0.14	0.21
		3	0.22	0.16
	2	1	13.46	17.76
		2	2.69	2.12
		3	1.01	1.32
	3	1	13.86	17.96
		2	1.14	5.05
		3	2.53	2.21
50	1	1	0.04	0.16
		2	0.00	0.03
		3	0.00	0.13
	2	1	0.35	0.00
		2	0.08	0.12
		3	0.07	0.03
	3	1	3.26	3.06
		2	0.21	0.23
		3	0.13	0.04

Table 2. Obtained Results for the Large-Sized Instances

$\|K\|$	$\bar{C}_L$	$\bar{D}_L$	$Gap_{Aff}\%$	$Gap_{Clos}\%$
10	1	1	6.91	22.34
		2	0.63	2.16
		3	1.99	1.33
	2	1	45.73	128.90
		2	4.99	26.81
		3	2.09	8.46
	3	1	22.32	146.58
		2	15.66	38.84
		3	3.56	22.36
50	1	1	1.00	3.40
		2	0.67	0.49
		3	0.43	0.81
	2	1	12.62	31.97
		2	0.81	4.09
		3	0.45	1.91
	3	1	18.81	63.51
		2	2.09	7.78
		3	0.89	3.44

When the results for the small-sized instances are analyzed, the proposed solution approach, which includes both indices, is capable of finding solutions close to the GUROBI solutions, with percentage gaps of less than 1% in most cases. On the other hand, high percentage gaps are observed for the problems with the lowest demand level. Similar status is observed for the large-sized instances. It is clear from Table 2 that low percentage gap values are observed at demand levels 2 or 3, while high percentage gap values are observed at demand level 1. Inasmuch as the percentage gap value increases even further when the number of product types is 10. In particular, the performance of the proposed algorithm with the closeness index is poor for problems with low demand levels. As a consequence, it should be stated that the proposed solution approach can produce efficient results on the generated instances across most problem configurations. However, large percentage gaps can be observed when demand is low in the TPP. Another finding from the computational studies is that the proposed solution approach, which uses the affordability index, outperforms the algorithm that uses the closeness index. In particular, for the small-sized instances, it is observed that f_{Aff} is better than f_{Clos} for 28 instances, while f_{Clos} is better than f_{Aff} for 11 instances. For the large-sized instances, f_{Aff} is better than f_{Clos} for 81 instances, while f_{Clos} is better than f_{Aff} for only 9 instances. As a result, the affordability index yields a more successful market selection for the TPP than the closeness index.

V. CONCLUSION

In this study, the traveling purchaser problem (TPP) is considered. Unlike classical approaches, a heuristic approach based on graph theory is developed to solve the problem. To date, there is no approach to TSP or TPP using topological graph indices. In this context, two new topological graph indices, which are called the affordability index and the closeness index, are introduced. As their names suggest, these two graph indices are primarily related to the prices of goods in the list and the distances between the markets and the depot. The solutions obtained using these two topological graph indices are compared with those obtained without graph theory in the literature, and the advantages of this graph-theoretic approach over existing methods are revealed. In future work, one may focus on several other topological graph indices, and further improvements may be possible.

ACKNOWLEDGEMENT

This study was funded by Bursa Uludag University Research Fund (Grant number GAP4-2025/2204).

REFERENCES

[1] H. Wiener, “Structural determination of paraffin boiling points,” Journal of the American Chemistry Society, vol. 69(1), pp. 17-20, 1947.
[2] I. Gutman, and N. Trinajstic, “Graph theory and molecular orbitals III. Total π-electron energy of alternant hydrocarbons,” Chemical Physics Letters., vol. 17, pp. 535-538, 1972.
[3] K. C. Das, A. Yurttas, M. Togan, I. N. Cangul, and A. S. Cevik, “The multiplicative Zagreb indices of graph operations,” Journal of Inequalities and Applications, vol. 90, 2013.
[4] B. Furtula, and I. Gutman, “A forgotten topological index,” Journal of Mathematical Chemistry, vol. 53(4), pp. 1184-1190, 2015.
[5] J. B. Liu, A. Bahadur, M. A. Malik, H. M. A. Siddiqui, and M. Imran, “Reformulated Zagreb indices of some derived graphs,” Mathematics, vol. 7(4), pp. 366, 2019.
[6] A. Yurttas, M. Togan, and I. N. Cangul, “Zagreb Indices of graphs with added edges,” Proceedings of the Jangjeon Mathematical Society, vol. 21(3), pp. 385-392, 2018.

[7] D. Manerba, R. Mansini, and J. Riera-Ledesma, "The traveling purchaser problem and its variants," European Journal of Operational Research, vol. 259(1), pp. 1-18, 2017.
[8] G. Laporte, J. Riera-Ledesma, and J.-J. Salazar-González, "A branch-and-cut algorithm for the undirected traveling purchaser problem," Operations Research, vol. 51(6), pp. 940-951, 2003.
[9] S. Kang, and Y. Ouyang, "The traveling purchaser problem with stochastic prices: Exact and approximate algorithms," European Journal of Operational Research, vol. 209(3), pp. 265-272, 2011.
[10] H. Cambazard, and B. Penz, "A constraint programming approach for the traveling purchaser problem," In M. Wallace (Ed.), Principles and Practice of Constraint Programming – CP 2004, Springer, pp. 735-749, 2012

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Machine Learning in Agile Project Management: A Bibliometric Analysis

Yadira Jazmín Pérez Castillo
Centro de Investigación en Computación, Instituto Politécnico Nacional,
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, Alcaldía Gustavo A. Madero, 07738, Ciudad de México, México
yperezc2022@cic.ipn.mx
ORCID: 0000-0001-9745-1964

Sandra Dinora Orantes Jiménez
Centro de Investigación en Computación, Instituto Politécnico Nacional,
Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, Alcaldía Gustavo A. Madero, 07738, Ciudad de México, México
dinora@cic.ipn.mx
ORCID: 0000-0002-8626-5936

Abstract
In recent years, the use of Machine Learning techniques in agile project management has sparked growing interest within the scientific community. However, a clear understanding of the main research lines, emerging trends, and gaps in this interdisciplinary area is still lacking. This paper presents a bibliometric analysis examining the evolution of scientific output on ML in agile contexts, based on publications indexed in databases such as Scopus, IEEE, and Web of Science. Bibliometric and machine learning techniques were used to identify the most productive authors, themes, and application areas. The results show sustained growth in the field in recent years, with emphasis on effort prediction, team performance evaluation, and automation of sprint tracking. Finally, future research opportunities are discussed, integrating ML with agile dynamics in multi-project and collaborative scenarios.

Keywords: Machine Learning, Agile Project Management, Bibliometric Analysis.

I. INTRODUCTION

Over the last decade, Agile Project Management (APM) has gained significant adoption in software development contexts, standing out for its iterative approach, adaptability to change, and focus on continuous value delivery. In parallel, Machine Learning (ML) techniques have emerged as key tools for data analysis and decision automation in various domains, including software engineering. For example, [1] explores the challenges of large-scale agile projects and the need for advanced predictive techniques. The convergence of agile approaches and ML algorithms has given rise to a new line of research focused on optimizing sprint tracking, effort estimation, risk prediction, and backlog prioritization, among other key processes. Several studies, such as [2], present extensive literature reviews of machine learning applications in agile project management up to 2022. Unlike previous systematic literature reviews, this work provides a bibliometric perspective on publication trends, collaboration patterns, thematic clustering, and emerging research directions from 2020 to 2025. This complementary approach enables the identification of research dynamics that cannot be observed through traditional qualitative reviews alone. Additionally, [3] reaffirms the current state of artificial intelligence in software engineering and its potential for integration with the discipline.

II. METHODOLOGY

To identify the main lines of research at the intersection of ML and APM, a bibliometric study was conducted using academic publications retrieved from databases such as Scopus, IEEE, and Web of Science. The analysis focused on the period from 2020 to 2024 to capture the most recent trends in this emerging field.

1. Data source and search criteria

The Scopus, IEEE, and Web of Science databases were used due to their multidisciplinary coverage and recognition as reliable sources for bibliometric analysis. The advanced search was constructed with the following logical string: Search String: *(TITLE-ABS-KEY ("agile project management") OR TITLE-ABS-KEY ("agile software development") AND (TITLE-ABS-KEY ("machine learning") OR TITLE-ABS-KEY ("artificial intelligence")) AND (PUBYEAR > 2019)*

This query was designed to retrieve publications whose title, abstract, or keywords explicitly referenced both agile methodologies and ML techniques.

2. Extraction and Preprocessing Process

The results were exported as CSV from the selected databases, including fields such as title, authors, affiliations, abstract, publication year, keywords, and number of citations. The data were cleaned and normalized to ensure consistency (e.g., unifying key terms and removing duplicates). A first manual verification was then conducted to confirm that the titles focused on ML applied to APM and to discard unrelated works.

3. Analysis Tools

The bibliometric analysis was performed using a combination of open-source tools, including Python and ML techniques, to fully exploit the information contained in the article metadata. The main technologies and processes used are described below.

Bibliometrix (R and Python): The initial phase employed the Bibliometrix library (R version 4.x), along with its Python counterpart, which allows the extraction of basic scientific productivity indicators. The data cleaning and preprocessing were done using Python 3.9 with the following libraries:

- **Pandas**: to load, clean, and transform the CSV file, eliminating duplicates and incomplete records.
- **scikit-learn** (sklearn): TfidfVectorizer was used to convert the combined text (title + keywords) into numerical vectors, and K-Means was used to segment the publications into thematic clusters automatically.
- **matplotlib** and **seaborn**: to plot temporal trends.
- **WordCloud**: to generate word clouds visualizing the frequency of key terms in the corpus.

III. RESULTS

1. Cleaning and Number of Records

The original CSV was loaded, containing 282 rows selected for analysis and 4 columns (Title, Authors, Year, and Keywords). After removing duplicates and entries without a title or keywords, 261 records remained for further analysis.

2. Publications per Year

The publication years were analyzed to observe the temporal evolution of research on ML applied to agile management, as shown in Figure 1. The year 2021 presents the highest publication peak, with approximately 72

articles. From 2020 to 2024, the results show an overall upward trend, despite a slight decrease in 2023 followed by a recovery in 2024.

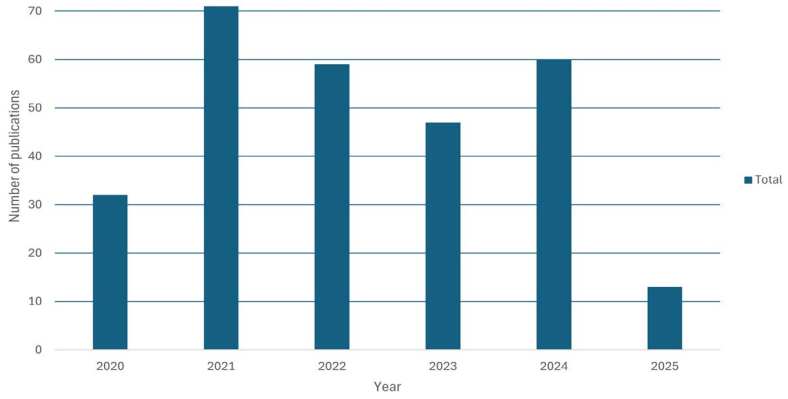


Figure. 1 Publications per Year

For 2025, only 13 studies were identified; however, this figure reflects only partial data collected up to the review's cutoff date, as the studies were conducted that same year and the publication cycle had not yet concluded.

3. Most Productive Authors

Author names were separated (using ";" or ","), and their publication counts were tallied, selecting the top 10 most frequent authors. Results are shown in Table 1. Four authors lead with three publications each. The rest of the top 10 have two publications. This shows that there is no dominant author, but rather several contributors with similar output.

Table 1. Productive Authors

Authors	Total of Publications
L. Chen	3
R. Ferdiana	3
B. Kumar	3
D. C. Dobhal	3
A. Egyed	2
C. S. González-González	2
Y. A. Korablev	2
L. A. Kutsenok	2
M. P. Uysal	2
C. Mayr-Dorn	2

4. Keyword Analysis

The author keywords were unified and extracted to obtain the frequency of the most repeated keywords (top 15), as shown in Figure 2.



Figure 2. Keyword WordCloud

Results indicate that “machine learning” leads by a wide margin (94 mentions). Secondary key themes include “artificial intelligence,” “agile,” and “project management.” The cloud also highlights terms like “software development,” “effort estimation,” “scrum,” and “natural language processing.”

5. Document Type

Document types were counted and shown in Figure 3, indicating that publications in this area tend slightly toward conferences over journals. However, some academic databases were inaccessible due to limitations, so including additional sources could enrich the analysis and more likely reveal additional document types.

6. Areas of applications

A brief description is provided below of how a clustering method was applied to the texts (Titles + Keywords) to identify application areas and classify the collected works. This is followed by Table 4, which presents the 5 most relevant articles as representative examples after the classification.

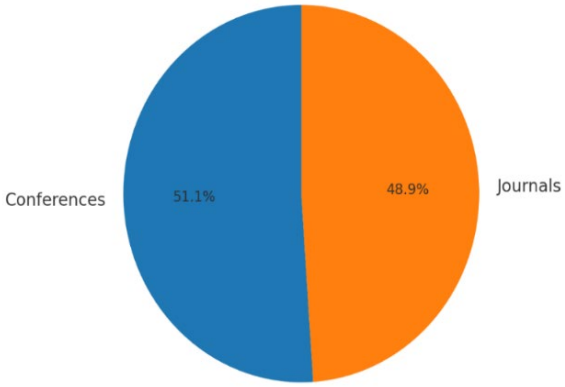


Fig. 3 Distribution of Document Types

The K-means clustering algorithm was selected because it is an efficient and well-established unsupervised clustering method for grouping documents based on their vector representations [4], [5]. By projecting each record (a combination of title and keywords) into a numerical feature space, K-Means enables the automatic identification of clusters of works that share similar vocabulary and thematic structures. Since the goal of the research was to objectively discover application areas without imposing predefined labels, K-Means provides a simple way to partition the dataset into k groups with high internal cohesion and external separation between

clusters. Moreover, its resulting centroids (mean vectors) facilitate the extraction of the most representative terms from each group, enabling manual assignment of descriptive names afterward. In short, K-Means was used because it combines computational efficiency with interpretability of results, and because its operation over TF-IDF vectors supports the detection of latent content patterns in the corpus of titles and keywords.

7. Text Preparation

Each record from the table (title and keywords) was combined into a single lower-case text field. Example:

- Title: "Risk Prediction in Agile Projects Using Bayesian Networks"
- Keywords: "bayesian network; risk management; Scrum"
- Combined text: "Risk prediction in agile projects using Bayesian networks, Bayesian network risk management, Scrum"

8. TF-IDF Vectorization

We used a TF-IDF vectorizer (removing terms that appeared in more than 80% of the documents and those appearing in fewer than 2 documents). This produced a sparse matrix of size ($n = 261$ *rows*) ($m = 3,$ 000 *columns*), where each column corresponds to a relevant term extracted from all texts.

9. Clustering with K-Means.

K-Means was executed with k = 10 (using a fixed seed for reproducibility), such that each of the 261 records was assigned to one of the 10 clusters (labeled 0–9). The value k = 10 was selected after several exploration experiments, seeking a balance between thematic granularity and interpretability of the resulting clusters. Lower values produced overly broad categories, while higher values generated fragmented clusters with limited semantic distinction.

10. Centroid Inspection

For each of the 10 clusters, the 10 highest-weight words (with the highest TF-IDF scores) from the centroid (average vector) were extracted. These "top words" help interpret the overarching theme of each cluster.

11. Assignment of Descriptive Labels

Based on the most representative terms identified in each centroid, a descriptive application-area label was manually assigned to each cluster. For instance, clusters characterized by terms such as "risk", "management", "assessment", and "Bayesian" were labeled as "Risk Management & Predictive Models".

12. Final results

Each document received an application area label (one of 10 names). The number of records in each cluster was counted to determine their distribution; the results are shown in Tables 2 and 3. The most-populated area is "Risk Management & Predictive Models" (75 papers), followed by "NLP & User Stories" (45) and "Other Topics" (36). "Effort Estimation and Cost Models" has the lowest share with 6 publications. Table 4 presents the following representative articles considered "most relevant," selected based on the presence of key terms.

Table 1. Discovered Application Area (Cluster Labels)

Cluster ID	Application Area	Description
0	Effort Estimation and Cost Models	Regression, Random Forest, COCOMO II + ML for predicting effort and cost.
1	Development Lifecycle & Agile Frameworks	Applying ML across the lifecycle (requirements, coding) and agile methods.
2	NLP & User Stories	NLP for classification, generation, and prioritization of user stories.
3	Defect Prediction & Quality Assurance	ML models to anticipate bugs and optimize test cases.
4	Risk Management & Predictive Models	Bayesian networks and decision trees for identifying and mitigating sprint risks.
5	Resource & Task Allocation	ML for recommending balanced task/resource assignments.
6	Hybrid Integration / Composite Frameworks	ML combined with Kanban/Scrum/DevOps pipelines.
7	Sentiment Analysis & Team Productivity	NLP to analyze team mood, feedback, and productivity.
8	Time Prediction / Scheduling	Time-series models and regression for forecasting sprint duration and delivery dates.
9	Other Topics (Various ML Applications)	Anomaly detection, task clustering, recommender systems.

Table 2. Publications Per Discovered Cluster

Application Area	Count
Risk Management & Predictive Models	75
NLP & User Stories	45
Other Topics	36
Lifecycle & Agile Frameworks	32
Defect Prediction & Quality Assurance	22
Sentiment & Team Productivity	15
Resource & Task Allocation	12
Time Prediction / Scheduling	10
Hybrid Integration / Composite Frameworks	8
Effort Estimation and Cost Models	6

Table 3. Top 5 Representative Articles

Title	Year
Sprint2Vec: A Deep Characterization of Sprints in Iterative Software Development [6]	2025
Using sentiment analysis to assess PMBOK knowledge areas' compatibility with agile [7].	2024
Sprint Management in Agile Approach: Progress and Velocity Evaluation Applying Machine Learning [8]	2024

K-Means and Feature Selection Mechanism to Improve Performance of Clustering User Stories in Agile Development [9]	2023
Prediction of Risk Percentage in Software Projects by Training Machine Learning Classifiers [10]	2021

13. Summary of Key Findings

The authors in [2] conducted a systematic review of machine learning applications in agile Project management, identifying trends, techniques, and research gaps in this field. They noted the growing interest in combining NLP techniques and deep learning models for tasks such as effort estimation and user story classification. This analysis complements the work of [3], who critically discuss the current state of artificial intelligence in software engineering and point out which areas, including agile methodologies, are more mature for adoption and which require further research. [11] proposed a ChatGPT-based approach for retrieving quality attributes in software development strategies, demonstrating that transformer-based language models can extract quality requirements (e.g., maintainability, usability) from user story descriptions. Similarly, [12] presented an NLP-based method for analyzing software requirements to improve effort estimation in agile environments, combining entity extraction and semantic analysis techniques to generate more accurate predictions. Reference [13] focused on privacy requirements engineering. They used NLP techniques to analyze user stories and detect privacy implications early, facilitating the integration of privacy controls into agile cycles. In the area of user story classification, [9] used K-Means with a feature selection mechanism to improve clustering performance of user stories in agile environments; their approach reduces the dimensionality of the textual space and groups stories more coherently, which aids in task assignment. Similarly, [3] proposed a classification method for product backlog items using machine learning (e.g., SVMs and decision trees) to automate the categorization of user stories by priority and complexity. [14] explored the implementation of ML in the Use Case Points (UCP) method, offering a 30-year retrospective analysis and showing how regression models and neural networks have evolved to better adjust complexity factors in UCP. Along the same lines, [15] described various predictive Models for project attributes (e.g., cost, duration) using traditional ML techniques (linear and multiparameter regression), emphasizing that using features derived from code metrics improves prediction accuracy. In [16] was demonstrated how an artificial neural network can enhance project duration prediction compared to classic statistical methods, by capturing nonlinear relationships between input variables. Regarding effort estimation in agile methodologies, [17] used decision trees and story points to model project effort and cost, concluding that decision trees provide more intuitive explanations and realistic results when historical metrics are available. In [18] applied k-Nearest Neighbors to estimate effort in agile projects, showing that a neighbor-based strategy allows for easy incorporation of similar past projects to improve predictions. [19] Compared linear regression, ridge regression, and logistic regression to forecast effort in agile projects, noting that ridge regression outperforms linear when multicollinearity is present among story point variables. [10] trained machine learning classifiers (SVMs, trees) to predict risk percentage in software projects, demonstrating that a classification approach can identify high-risk projects with over 80% accuracy. In [20] investigated the challenges of cost estimation in agile environments, noting that the lack of standardized metrics and high variability in requirements make statistical modeling difficult. They proposed combining historical data with machine-learning-based heuristic approaches. In [1] the authors discussed the challenges of agile projects, arguing that organizational complexity and the coordination among multiple teams demand sophisticated forecasting methods and real-time data analytics. [6] presented *Sprint2Vec*, a deep model that characterizes sprints in iterative Development through embeddings that capture attributes such as speed, collaboration, and story completion rate, facilitating sprint comparisons and identification of improvement patterns. In parallel, [8] applied ML techniques to evaluate sprint progress and velocity, proposing automated

metrics to predict future performance based on historical data. [21] addressed the determination of ideal team size in agile practices using learning-based methods, incorporating variables such as task complexity, prior experience, and cognitive load to recommend configurations that maximize productivity. [22] used deep learning to classify the functional size of agile projects under the COSMIC standard, demonstrating that convolutional neural networks can extract patterns directly from requirements descriptions. [7] Applied sentiment analysis to assess the compatibility of PMBOK knowledge areas with agile methodology, suggesting that some areas (e.g., communications management) show greater alignment, while others (e.g., risk management) require adjustments in terminology and practices. In a complementary approach, [23] developed Emotimonitor. This Trello plugin captures and monitors the emotions of agile teams in real time using text analysis techniques and brief surveys that help anticipate potential morale issues or conflicts. In [24] validated agile methodology as an ideal process using the Fuzzy-TOPSIS method, concluding that flexibility and responsiveness rank it above traditional models when fuzzy performance metrics are applied. [25] proposed ML models to predict the adoption of agile methodologies by organizations, based on structural and cultural characteristics of companies; their work showed that decision trees and Random Forest can identify with 85 accuracy which organizations are likely to adopt agile practices. Similarly, [26] presented a method for evaluating the success of agile projects using ML, combining time, cost, and customer satisfaction metrics to generate scores that predict the likelihood of success. [27] designed a task allocation approach for distributed agile environments using ML, where a clustering model suggests the best distribution of user stories based on skills and availability, reducing coordination overhead. [24] used Fuzzy-TOPSIS to validate the suitability of agile methodology in software projects, finding that flexibility and fast feedback are the main factors making it preferable to traditional waterfall methods. This finding complements the study by [25], which highlighted that organizational culture plays a decisive role in the successful adoption of Agile. In summary, recent literature combines NLP and ML tools to address key challenges in agile environments: from the automatic extraction of requirements and quality attributes [11], [13] to the prediction of project metrics such as effort, cost, and duration, [14], [15], [16], [17], [18], [19], [25] Advanced techniques are also being explored to characterize sprints [6] [22] [8] and monitor human aspects such as emotions and framework compatibility [7], [23]. Large-scale studies highlight the need for robust models that incorporate cultural and organizational metrics [24], [25], [1], [26] while others focus on optimizing specific tasks like automatic task allocation in distributed teams [27] or determining ideal team sizes [21]. These contributions show a clear trend toward integrating AI and Big Data to enhance efficiency, quality, and predictability in agile projects.

IV. DISCUSSION

Between 2020 and 2024, publications on ML applied to APM grew from 32 articles in 2020 to 71 in 2021, then stabilized in 2022 and 2023 (59 and 47, respectively), before rising again to 60 in 2024. This growth reflects sustained interest in combining both disciplines. Author contributions are distributed; no dominant figure exists, with most authors contributing 2–3 papers, indicating a collaborative ecosystem. In terms of dissemination, 51.1% of publications appear in conferences and 48.9% in journals, showing a balanced use of both forums. Keyword analysis shows that the core topics are “machine learning,” “agile,” and “project management.” Subfields like “effort estimation” and “natural language processing” shape specific research directions, ranging from regression models to user story classification. Using TF-IDF and K-Means clustering, ten distinct application areas were identified. The most frequent were:

- Risk Management & Predictive Models (28.7%)
- NLP & User Stories (17.2%)
- Other Topics (13.8%)
- Lifecycle & Agile Frameworks (12.3%)
- Defect Prediction & QA (8.5%)

- Other smaller categories included sentiment analysis, resource allocation, time prediction, and ML-DevOps hybrid environments.

These findings offer a detailed view of how ML is applied in agile contexts, highlighting both consolidated focus areas (e.g., risk management, NLP for requirements) and emerging fields. The bibliometric results allow us to reflect on the current research landscape on the use of ML in APM. The remarkable increase in scientific output over the last five years indicates strong interest from both academia and industry in integrating AI techniques with agile frameworks such as Scrum, Kanban, and XP. A notable aspect is the thematic dispersion found in keyword co-occurrence networks. While core terms are prevalent, many peripheral nodes suggest the field is still in an exploratory phase with various theoretical and practical approaches in development. This diversity signals that the field remains open to new frameworks, tools, and methodologies.

V. LIMITATIONS

This study presents several limitations. First, only publications indexed in Scopus, IEEE, and Web of Science were considered, potentially excluding relevant studies in other scientific databases. Second, the clustering results depend on the selected search string and the value of k used in the K-Means algorithm, meaning that different parameter configurations can yield alternative thematic groupings. Third, assigning descriptive labels to clusters involved manually interpreting the most representative terms, which may introduce some subjectivity. Finally, the analysis was limited to bibliographic metadata, such as titles, keywords, and publication information, and did not include a full-text analysis of the retrieved studies.

VI. CONCLUSIONS

This study has shown that the combination of Machine Learning with agile methodologies has gained increasing relevance in recent years, as evidenced by the significant rise in publications since 2021. The bibliometric analysis revealed a highly collaborative author ecosystem, with no hegemonic researcher, suggesting convergence among multiple teams working on ML solutions to optimize agile processes. The most recurrent keywords, such as “machine learning,” “agile,” and “project management,” highlight the central research focus, while terms like “effort estimation” and “natural language processing” indicate specialization into subdomains. The use of TF-IDF and K-Means clustering helped identify ten differentiated application areas, with “Risk Management & Predictive Models” and “NLP & User Stories” being the most prominent. This expands the traditional focus on effort estimation and user stories, since nearly half of the papers fall under “Other Topics,” covering everything from defect prediction to ML-DevOps hybrid support. The near-equal distribution between conferences (51%) and journals (49%) suggests that the academic community values both the immediacy of conferences and the rigor of journals to share findings in this interdisciplinary space. Together, the results provide a comprehensive overview of how ML techniques are applied to agile tasks, whether for predicting risks, classifying user stories, forecasting timelines, or recommending resource assignments, and underline the many still-exploratory opportunities. These conclusions validate the growing synergy between ML and agile management, while also revealing key gaps, such as the limited application of ML to sprint tracking. The findings can serve as a foundation for future studies aiming to standardize methods, build shared repositories, and validate models in real-world settings.

VII. FUTURE WORK

Future work should focus on developing ML models that integrate seamlessly with daily agile practices, enabling real-time tracking of project progress and health. For instance, using recurrent neural networks or transformers specialized in time series, it would be possible to accurately predict sprint progress based on historical card

updates and comments, generating automatic alerts when delays are likely. Additionally, a dynamic dashboard that combines key metrics (burn-down, velocity, defect rate) with classification and clustering algorithms could proactively suggest task reassignment or priority reevaluation, and even anticipate bottlenecks. Moreover, by leveraging natural language processing techniques, a virtual assistant (Agile Coach) could analyze text from stand-ups, pull requests, and feedback to detect blockers, negative sentiment, or team overload and offer precise recommendations without disrupting workflow. In parallel, a risk early-warning system could be built that, fed by sprint indicators (blocked stories, ticket latency, software quality metrics), computes a daily "risk score" for each critical item. When certain thresholds are exceeded, the tool would notify the Scrum Master to trigger replanning or retrospective meetings. A continuous quality tool could also be implemented to estimate the likelihood of introducing defects in new stories and to show a quality semaphore on the Kanban board. Finally, applying temporal clustering to sprint productivity metrics would allow the identification of patterns of "calm sprints" or "volatile sprints" and suggest adjustments to Work in Progress (WIP) or story refinement processes to optimize team performance. With these developments, integrating ML into agile tracking would not only anticipate problems but also provide predictive and automated solutions to sustain the iterative flow and continuous improvement.

ACKNOWLEDGEMENT

The authors would like to thank the *Instituto Politécnico Nacional and the Centro de Investigación* en Computación for their academic and institutional support. The authors also thank the participants who contributed to the evaluation of the proposed approach.

REFERENCES

[1] H. Saeeda, M. O. Ahmad, and T. Gustavsson, "Challenges in large-scale agile software development projects," in *Proc. 38th ACM/SIGAPP Symp. Appl. Comput. (SAC '23)*, Tallinn, Estonia, 2023, pp. 1030–1037, doi: 10.1145/3555776.3577662.

[2] Y. J. Pérez-Castillo, S. D. Orantes-Jiménez, and P. O. Letelier-Torres, "A systematic literature review on machine learning applications for agile project management," *Ingeniería Investigación y Tecnología*, vol. 25, no. 3, pp. 1–11, 2024, doi: 10.22201/fi. 25940732e.2024.25.3.017.

[3] A. Mashkoor, T. Menzies, A. Egyed, and R. Ramler, "Artificial intelligence and software engineering: Are we ready?" *Computer*, vol. 55, no. 3, pp. 24–28, Mar. 2022, doi: 10.1109/MC.2022.3144805.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, Berkeley, CA, USA, 1967, vol. 1, pp. 281–297.

[5] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. KDD Workshop Text Mining*, Boston, MA, USA, 2000, pp. 109–110. DOI not available.

[6] M. Choetkiertikul, P. Banyongrakkul, C. Ragkhitwetsagul, S. Tuarob, H. K. Dam, and T. Sunetnanta, "Sprint2Vec: A deep characterization of sprints in iterative software development," *IEEE Trans. Softw. Eng.*, vol. 51, no. 1, pp. 220–242, Jan. 2025, doi: 10.1109/TSE.2024.3509016.

[7] I. David and R. Gelbard, "Using sentiment analysis to assess PMBOK knowledge areas' compatibility with agile methodology," *Procedia Comput. Sci.*, vol. 239, pp. 381–395, 2024, doi: 10.1016/j.procs.2024.06.185.

[8] Y. J. Pérez-Castillo, S. D. Orantes-Jiménez, and P. O. Letelier-Torres, "Sprint management in agile approach: Progress and velocity evaluation applying machine learning," *Information*, vol. 15, no. 11, Art. no. 726, Nov. 2024, doi: 10.3390/info15110726.

[9] R. A. Putri, U. L. Yuhana, and T. C. Amri, "K-means and feature selection mechanism to improve performance of clustering user stories in agile development," in *Proc. Int. Conf. Modeling E-Information Research, Artificial Learning and Digital Applications (ICMERALDA)*, 2023, pp. 39–43, doi: 10.1109/ICMERALDA60125.2023.10458165.

[10] G. Panneerselvam and S. Sankaranarayanan, "Prediction of risk percentage in software projects by training machine learning classifiers," *Comput. Electr. Eng.*, vol. 94, Art. no. 107362, Sep. 2021, doi: 10.1016/j.compeleceng.2021.107362.

[11] A. Aamana, Q. U., and S. U. Nisa, "Beyond agile: NLP-driven quality attributes retrieval using ChatGPT in software development strategies," in *Proc. Int. Conf. Engineering and Computing Technologies (ICECT)*, 2024.

[12] T. Catak, P. O. Durdu, and S. I. Omurca, "Enhancing agile effort estimation: An NLP approach for software requirements analysis," in *Proc. Int. Congr. Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2024, pp. 1–8, doi: 10.1109/HORA61326.2024.10550870.

[13] G. B. Herwanto, G. Quirchmayr, and A. M. Tjoa, "Leveraging NLP techniques for privacy requirements engineering in user stories," *IEEE Access*, vol. 12, pp. 22167–22189, 2024, doi: 10.1109/ACCESS.2024.3364533.

[14] R. S. Dewi, R. Sarno, and E. S. Astuti, "Machine learning implementation in use case points method: A 30-year perspective," in *Proc. Int. Conf. Information Technology Systems and Innovation (ICITSI)*, 2024.

[15] C.-L. Fan, "Prediction models for project attributes using machine learning," in *Advances in Intelligent Systems and Computing*, Lecture Notes in Networks and Systems, vol. 308. Cham, Switzerland: Springer, 2022, pp. 238–245, doi: 10.1007/978-3-030-85577-2_28.

[16] I. Lishner and A. Shtub, "Using an artificial neural network for improving the prediction of project duration," *Mathematics*, vol. 10, no. 22, Art. no. 4189, Nov. 2022, doi: 10.3390/math10224189.

[17] E. Rodríguez-Sánchez, E. F. Vázquez-Santacruz, and H. Cervantes-Maceda, "Effort and cost estimation using decision tree techniques and story points in agile software development," *Mathematics*, vol. 11, no. 6, Art. no. 1477, Mar. 2023, doi: 10.3390/math11061477.

[18] E. R. Sánchez, H. C. Maceda, and E. V. Santacruz, "Software effort estimation for agile software development using a strategy based on the k-nearest neighbors' algorithm," in *Proc. IEEE Mexican Int. Conf. Computer Science (ENC)*, 2022, doi: 10.1109/ENC56672.2022.9882947.

[19] M. Vyas and N. Hemrajani, "Predicting effort of agile software projects using linear regression, ridge regression, and logistic regression," *Int. J. Tech. Phys. Probl. Eng.*, vol. 13, no. 2, pp. 14–19, 2021.

[20] S. A. Saeed, J. A. Khan, S. Naeem, and S.-U.-R. Khan, "An empirical investigation on cost estimation challenges in agile software development context," in *Proc. Int. Conf. Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, 2021, pp. 188–193, doi: 10.1109/FIT53504.2021.00043.

[21] R. Olivares, R. Noel, S. M. Guzmán, D. Miranda, and R. Muñoz, "Intelligent learning-based methods for determining the ideal team size in agile practices," *Biomimetics*, vol. 9, no. 5, Art. no. 292, May 2024, doi: 10.3390/biomimetics9050292.

[22] Y. S. Molla, S. T. Yimer, and E. Alemneh, "COSMIC-functional size classification of agile software development: Deep learning approach," in *Proc. Int. Conf. Information and Communication Technology for Development for Africa (ICT4DA)*, 2023, pp. 155–159, doi: 10.1109/ICT4DA59526.2023.10302232.

[23] M.-A. A. El-Migid, D. Cai, T. Niven, J. Vo, K. Madampe, J. Grundy, and R. Hoda, "Emotimonitor: A Trello power-up to capture and monitor emotions of agile teams," *J. Syst. Softw.*, vol. 186, Art. no. 111206, Apr. 2022, doi: 10.1016/j.jss.2021.111206.

[24] N. Govil and A. Sharma, "Validation of agile methodology as an ideal software development process using the fuzzy-TOPSIS method," *Adv. Eng. Softw.*, vol. 168, Art. no. 103125, Jun. 2022, doi: 10.1016/j.advengsoft.2022.103125.

[25] R. Hanslo and M. Tanner, "Machine learning models to predict agile methodology adoption," in *Proc. Federated Conf. Computer Science and Information Systems (FedCSIS)*, 2020, pp. 857–865, doi: 10.15439/2020F214.

[26] D. Veido, B. Misnevs, and A. Plotkin, "The method of agile project success evaluation using machine learning," in *Reliability and Statistics in Transportation and Communication*, Lecture Notes in Networks and Systems, vol. 117. Cham, Switzerland: Springer, 2020, pp. 478–487, doi: 10.1007/978-3-030-44610-9_47.

[27] P. William, P. Kumar, G. S. Chhabra, and K. Vengatesan, "Task allocation in distributed agile software development using machine learning approach," in *Proc. Int. Conf. Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, 2021, pp. 168–172, doi: 10.1109/CENTCON52345.2021.9688114.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Real-Time Helmet Detection System with Monitoring and Alert for Restricted Areas

Mohammad Mohatram
Electrical and Electronics Engineering
Global College of Engineering and Technology
Muscat, Oman
m.mohatram@gcet.edu.om
ORCID:0000-0002-5459-5631

Javeed Hussain Shaik
Electrical and Electronics Engineering
Global College of Engineering and Technology
Muscat, Oman
s.javeedhussain@gcet.edu.om
ORCID:0000-0002-2696-1163

Asma Abdallah Al Reesi
Electrical and Electronics Engineering
Global College of Engineering and Technology,
Muscat, Oman
202011181@gcet.edu.om

Muhammad Irfan Memon
Electrical and Electronics Engineering
Global College of Engineering and Technology
Muscat, Oman
m.memon@gcet.edu.om
ORCID:0000-0003-4194-3657

Abstract
Occupational safety enforcement in industrial and construction environments remains a persistent global challenge, particularly in restricted areas where helmet compliance is mandatory. This paper presents a comprehensive real-time helmet detection and monitoring system that integrates YOLOv8-based deep learning with Arduino-controlled hardware alerts and a Streamlit dashboard for continuous compliance tracking. The system pipeline encompasses dataset construction and annotation, model fine-tuning, temporal filtering to stabilize detection, serial communication with embedded hardware (LED, buzzer, LCD), and CSV-based event logging with visual evidence. Experimental evaluation on a custom dataset demonstrates an mAP@0.5 of 0.95, a precision of 0.91, and a recall of 0.92, with alarm activation latency below 1 second. The dashboard reports 75% helmet compliance across monitored sessions. These results confirm the system's suitability for real-world deployment in low-resource industrial settings, providing both immediate safety enforcement and long-term compliance auditing.

Keywords: helmet detection, YOLOv8, computer vision, occupational safety, real-time monitoring, dashboard monitoring, restricted zones, temporal filtering, embedded systems.

I. INTRODUCTION

Enforcing occupational safety regulations in construction and industrial environments is a critical global priority. Mandatory helmet use in restricted zones, where workers are exposed to hazards such as falling objects, moving machinery, and hazardous materials, represents one of the most fundamental preventive measures. Despite stringent regulatory frameworks, compliance monitoring typically relies on manual supervision, which is inherently susceptible to human error, observer fatigue, inconsistencies, and coverage gaps during extended shifts. Recent advances in deep learning and computer vision have created compelling opportunities to automate personal protective equipment (PPE) compliance monitoring. Object detection frameworks, including Faster R-CNN [1], SSD, and the YOLO family, have demonstrated high accuracy in identifying safety equipment across diverse conditions. However, most published work focuses exclusively on detection accuracy, leaving the critical engineering challenge of real-world integration, such as hardware alerts, dashboard visualization, data logging, and compliance reporting, largely unaddressed. This paper presents a complete, end-to-end real-time helmet



detection and compliance monitoring system designed for deployment in restricted industrial areas. The system combines YOLOv8n, an advanced single-stage object detector, with an Arduino Uno-based alarm subsystem (LED indicator, buzzer, LCD), a temporal-filtering decision layer to stabilize frame-level predictions, and a Streamlit-based monitoring dashboard for live visualization and historical compliance analysis. The principal contributions of this work are threefold:

i) Development of a robust, annotated custom dataset and YOLOv8 training pipeline for two-class helmet detection, achieving mAP@0.5 of 0.95.
ii) Hardware integration of detection outputs with Arduino-controlled visual and auditory alarms via Python serial communication, enabling alarm activation in under one second.
iii) Implementation of temporal filtering to stabilize detections and a Streamlit dashboard for real-time monitoring and compliance auditing with CSV-based event logging.

By bridging the gap between research-grade detection accuracy and practical deployment in resource-constrained industrial settings, this system offers a scalable and low-cost framework for occupational safety monitoring.

II. LITERATURE REVIEW

Helmet detection research has evolved substantially over the past decade, progressing from traditional image processing methods toward deep learning frameworks that deliver real-time performance with practical accuracy. Early CNN-based approaches demonstrated the feasibility of binary helmet classification in still images but struggled with cluttered backgrounds and required separate object localization steps. Mercado et al. [2] reported moderate accuracy using convolutional classifiers for motorcycle helmet detection, establishing a benchmark for subsequent single-stage approaches. Chen et al. extended this work using Faster R-CNN, achieving improved mean average precision for construction site monitoring, but at computational costs that are incompatible with real-time embedded deployment. The YOLO framework has emerged as the dominant paradigm for real-time safety equipment detection, balancing accuracy and inference speed through single-stage processing. Wang et al. [3] demonstrated real-time helmet detection on construction sites using CSYOLOv3, achieving average precision exceeding 90% and leveraging focal loss strategies to address class imbalance [4] in live video streams. Liu and Hao [5] further improved detection performance using a modified YOLOv5 architecture for general PPE monitoring. Most recently, Zuo et al. [6] proposed an enhanced YOLOv8n three-module architecture with improved lightweight performance, while Wu et al. [7] introduced YOLOv8-CGS for construction site safety monitoring, confirming YOLOv8 as the current state-of-the-art for this application class. Despite these advances in detection, significant gaps remain in end-to-end system integration. Yadav et al. [8] proposed a deep learning approach for helmet detection with fine-grained generation for traffic surveillance, demonstrating classification performance but lacking integration with hardware alarms. Transformer-based architectures [9] have yet to be integrated into deployed systems. Similarly, smart monitoring systems reviewed in [10] typically require high-end GPU infrastructure, limiting their applicability in low-cost industrial contexts. Data preparation methodology is consistently identified as a critical factor in detection system performance. Standard annotation tools such as LabelImg [11] and CVAT are widely used for bounding-box labeling in YOLO-compatible formats. Data augmentation strategies like rotation, scaling, and photometric adjustments are employed to improve model generalization across diverse lighting conditions and viewpoints [12]. From this review, three specific gaps motivate the present work:

i) the absence of low-cost, integrated systems combining real-time YOLO-based detection with embedded hardware alarms;

ii) limited attention to detection stability under dynamic conditions, addressed here through temporal filtering; and
iii) the lack of a compliance dashboard and logging systems for long-term safety auditing in restricted industrial zones.

III. DESIGN AND IMPLEMENTATION

The proposed system was developed using a structured pipeline that integrates dataset preparation, model training, real-time detection, decision logic, hardware alarm control, data logging, and dashboard visualization. Each component was designed to ensure accurate detection, sub-second response, and comprehensive compliance monitoring. The system architecture is shown in Figure 1.

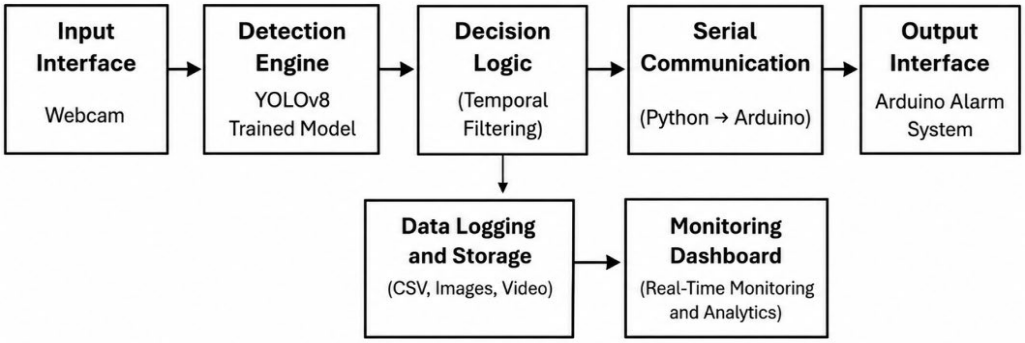


Figure 1. Architecture of the proposed real-time helmet monitoring and detection system

1. Dataset Collection and Annotation

A custom dataset was constructed by combining publicly available helmet detection repositories with supplementary images from diverse industrial and construction-site contexts. The dataset encompasses two object classes: helmet and head (bare, no helmet). Images were curated to represent challenging real-world conditions, for instance, varied head orientations (frontal, lateral, oblique), multiple workers per frame, both indoor and outdoor environments, and a range of ambient lighting conditions from direct sunlight to artificial indoor illumination, to promote generalization. Manual annotation was performed in YOLO-compatible format using LabelImg, with bounding boxes defined by normalized center coordinates and dimensions:

<class_id> <x_center> <y_center> <width> <height>

where class ID 0 denotes head (no helmet) and class ID 1 denotes helmet. Table 1 presents the data characteristics.

TABLE 1. Dataset Characteristics

Metric	Description
Total number of images	1,250 images
Total number of bounding boxes	3,400 annotations

Class distribution	Helmet: X images / Y boxes; Head (no helmet): X images / Y boxes
Environmental variability	Indoor (40%), outdoor (60%); daylight (70%), artificial/low light (30%); single worker (50%), multiple workers (50%)
Train/val/test split	70/15/15

2. Model Training

YOLOv8n was selected as the detection backbone for its balance of inference speed and detection accuracy in real-time safety monitoring. A pre-trained YOLOv8n model was fine-tuned using the Ultralytics Python framework on a custom dataset with an input resolution of 640 × 640 pixels for 30 epochs. The model was trained to detect two classes, helmet and head, using the AdamW optimizer with a cosine learning-rate schedule and focal loss to address class imbalance. Data augmentation techniques, including mosaic, random flipping, and HSV adjustment, were also applied to improve model generalization. Training was conducted on a GPU-enabled workstation. Performance metrics were monitored per epoch, including box loss, classification loss, precision, recall, and mAP at IoU thresholds of 0.5 and 0.5:0.95.

3. Real-Time Camera Integration and Decision Logic

The trained YOLOv8n model was integrated with a standard USB webcam via OpenCV, enabling continuous frame-by-frame inference. Detected objects are annotated with bounding boxes and class labels overlaid on the live video feed, providing continuous visual feedback for monitoring personnel.

To address frame-level prediction instability, a temporal filtering layer was implemented to complement attention-based stability techniques. To address frame-level prediction instability caused by sensor noise, sudden movements, and momentary occlusions, a temporal filtering layer was implemented. Rather than acting on single-frame detections, the system aggregates detection results across a configurable time window. The temporal window size was set to 10 frames. A SAFE state is declared if ≥7 of the last 10 frames detect a helmet. A WARNING state is declared if ≥7 of the last 10 frames detect no helmet. A hysteresis threshold of 2 frames prevents state oscillation. This approach substantially reduces false-positive alarms caused by transient misclassifications, improving system reliability in dynamic industrial environments. The detection pipeline is illustrated in Figure 2.

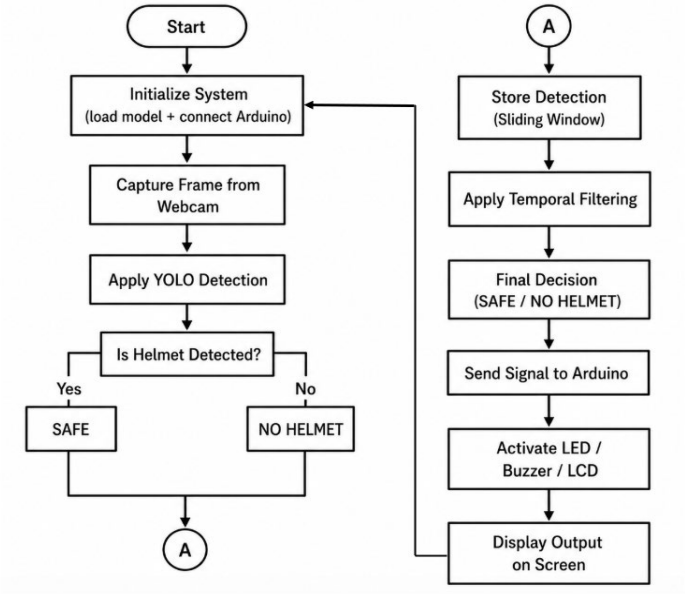
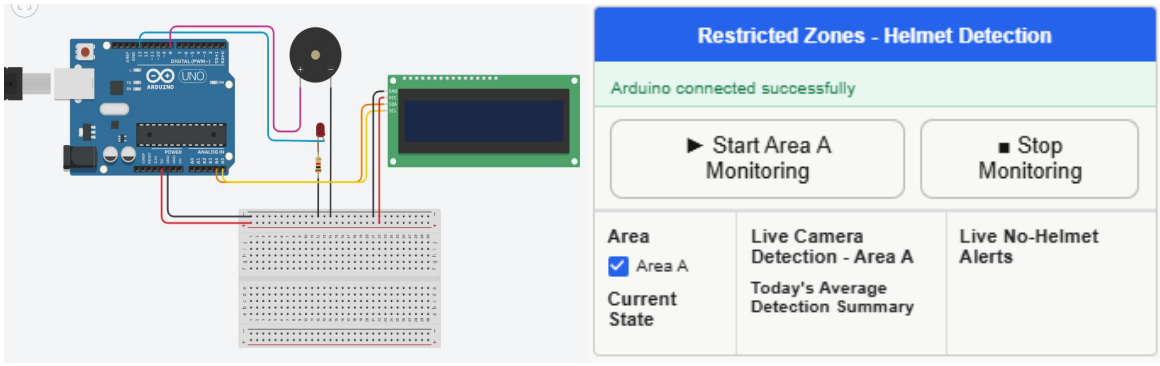


Figure 2. Real-time detection pipeline with temporal filtering decision logic

4. Alarm Subsystem

An Arduino Uno microcontroller was used to implement the physical alarm subsystem, consisting of green/red LED indicators for visual status signaling, a piezoelectric buzzer for audible warnings, and a 16×2 LCD for displaying the system state. The alarm circuit was initially validated through Tinkercad simulation before physical assembly. Python communicates with the Arduino via USB serial at 9600 baud, transmitting signals based on the temporal filtering output. In the SAFE state, the green LED remains on, the buzzer is deactivated, and the LCDs display "SAFE." In the WARNING state, the LED switches to red, the buzzer activates, and the LCDs "No Helmet." The hardware schematic and dashboard interface are shown in Figure 3. A Streamlit-based monitoring dashboard was developed to provide real-time system monitoring and historical compliance analysis. The dashboard displays a live camera feed with detection bounding boxes, the current system state (SAFE/WARNING), no-helmet alerts with timestamps and confidence scores, daily helmet compliance percentages, and stored video and image evidence. Figure 3(b) presents the Streamlit dashboard, including the live feed, system status, compliance statistics, and alert log.



(a) (b)

Figure 3. Hardware schematic and the screenshot of the dashboard interface

IV. RESULTS AND DISCUSSION

1. Training and Validation Performance

The YOLOv8n model exhibited consistent convergence over 30 training epochs. Box loss decreased from 1.41 to 1.13, reflecting improved helmet-head localization accuracy. Classification loss declined steadily, indicating effective learning of the two-class discrimination task. Training curves are presented in Figure 4.

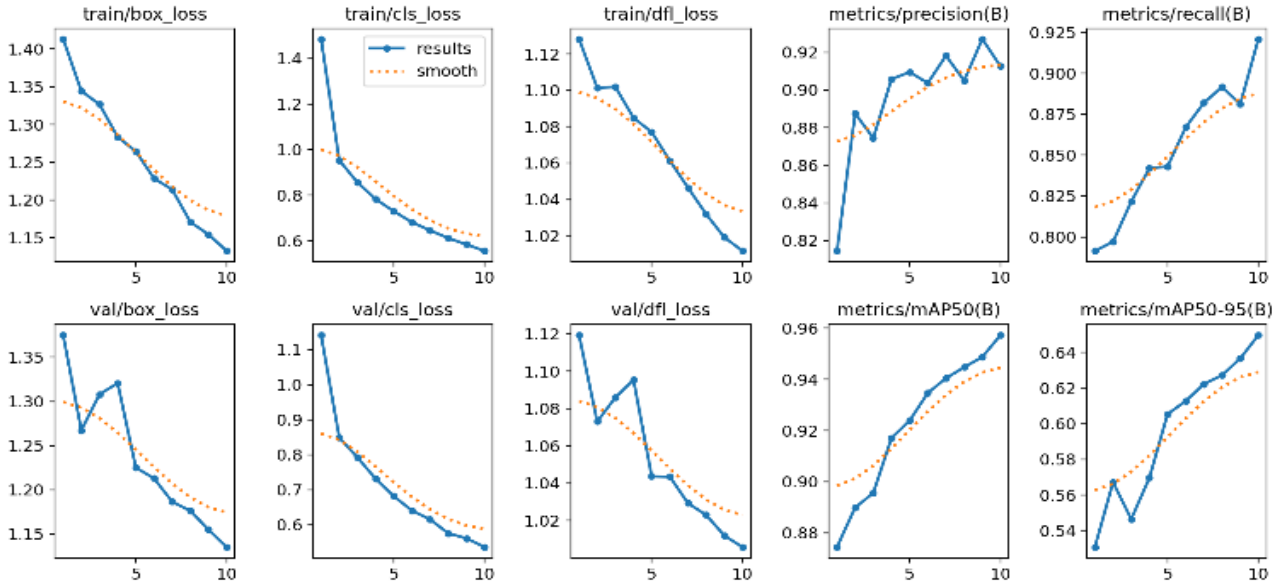


Figure 4. Training curves showing loss reduction and performance metric improvement across epochs

Final validation metrics are summarized in Table 2. Precision improved from 0.81 (epoch 1) to 0.91 (epoch 30), and recall increased from 0.79 to 0.92. mAP@0.5 reached 0.95, confirming high detection reliability at standard IoU thresholds. The mAP@0.5:0.95 of 0.65 reflects reduced performance at stricter localization thresholds, which is characteristic of real-world datasets with background clutter and partial occlusion.

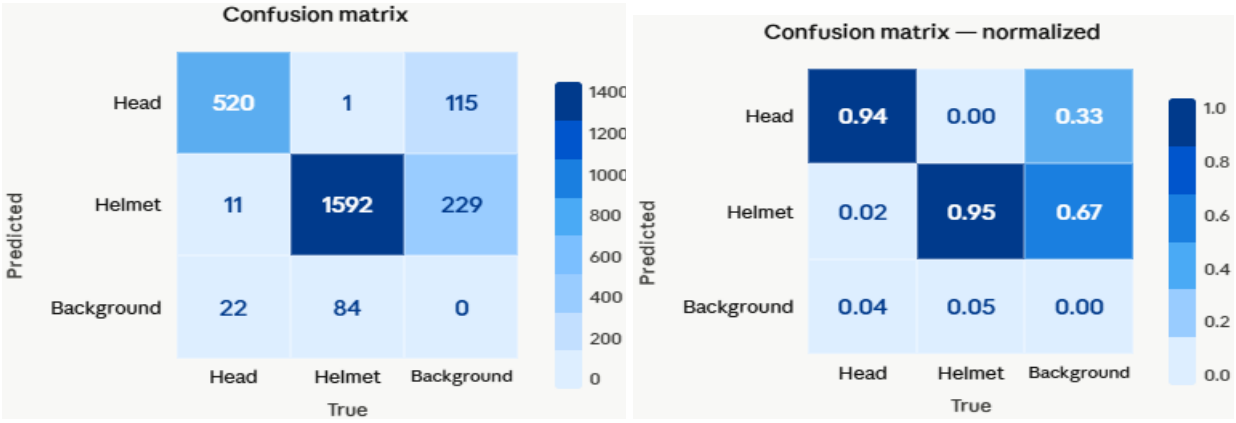
TABLE 2. YOLOv8n Training and Validation Performance Metrics

Metric	Epoch 1	Epoch 30
Precision	0.81	0.91
Recall	0.79	0.92
mAP@0.5	—	0.95
mAP@0.5:0.95	—	0.65
Box Loss	1.41	1.13

2. Detection Accuracy and Confusion Matrix

Evaluation on the held-out test set confirmed strong classification performance. The model achieved 95% classification accuracy for the helmet class and 94% for the head (no helmet) class. Background misclassification occurred in 5% of cases, reflecting the inherent challenge of cluttered industrial environments. Confusion matrices (both raw and normalized) are shown in Figure 5.

The temporal filtering layer further reduced false-positive alarm activations in dynamic scenarios, improving the operational reliability of the alarm subsystem beyond what raw per-frame accuracy metrics would suggest.



(a) Raw (b) Normalized

Figure 5. Confusion matrices showing class-wise prediction accuracy

3. Real-Time Detection and System Latency

Real-time system performance was verified through continuous webcam deployment. As detailed in Table 3, detection logs from representative monitoring sessions captured multiple no-helmet violations with confidence scores consistently ranging from 0.81 to 0.87. End-to-end inference and alarm activation latency were measured at under 200 ms per frame, enabling near-instantaneous safety enforcement. The Arduino alarm subsystem responded consistently within one second of violation detection in all tested scenarios. Sample real-time detection outputs with bounding box overlays are shown in Figure 6.

TABLE 3. Sample Detection Log with Alert Status

Timestamp	Class	Confidence	Area	Alarm
08:32:14	No Helmet	0.87	Area A	ON
08:33:02	Helmet	0.93	Area A	OFF
09:15:47	No Helmet	0.83	Area A	ON
09:16:01	No Helmet	0.81	Area A	ON
10:04:29	Helmet	0.95	Area A	OFF



Figure 6. Real-time detection outputs with bounding box overlays for helmet and head classes

4. Dashboard and Compliance Reporting

The Streamlit dashboard provided continuous real-time visualization and accumulated compliance statistics across monitored sessions. Daily compliance reports indicated 75% helmet usage and 25% noncompliance events. Recorded violations were stored with timestamps, confidence scores, and linked video/image evidence, enabling structured compliance auditing for regulatory purposes. This capability allows safety personnel to identify trends in noncompliance over time, flag high-risk periods (e.g., shift start and end transitions), and maintain documentary evidence for both internal reporting and external regulatory inspections. The dashboard interface is shown in Figure 7.

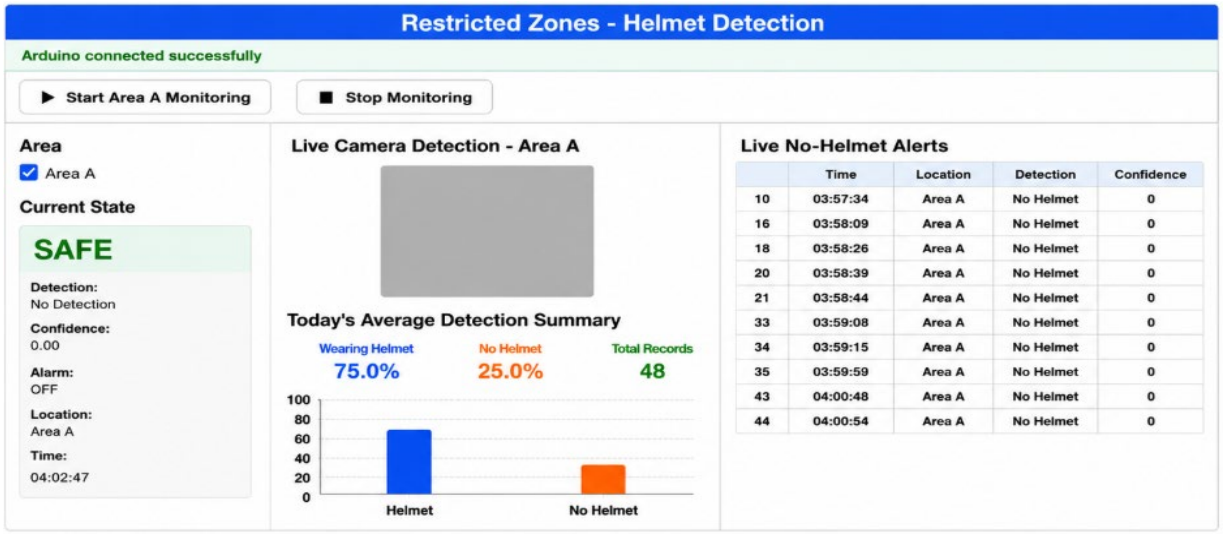


Figure 7. Dashboard interface showing live camera feed, system state, and compliance statistics

5. Hardware Integration Results

The Arduino alarm subsystem demonstrated reliable performance across all tested operational states. In SAFE mode, the LCD displayed “SAFE,” the LED remained green, and the buzzer was deactivated. In WARNING mode, the LCD displayed “No Helmet,” the LED turned red, and the buzzer activated within 1 second of detection confirmation. Hardware performance was consistent throughout extended testing sessions, with no observed hardware failures. The assembled hardware is presented in Figure 8.

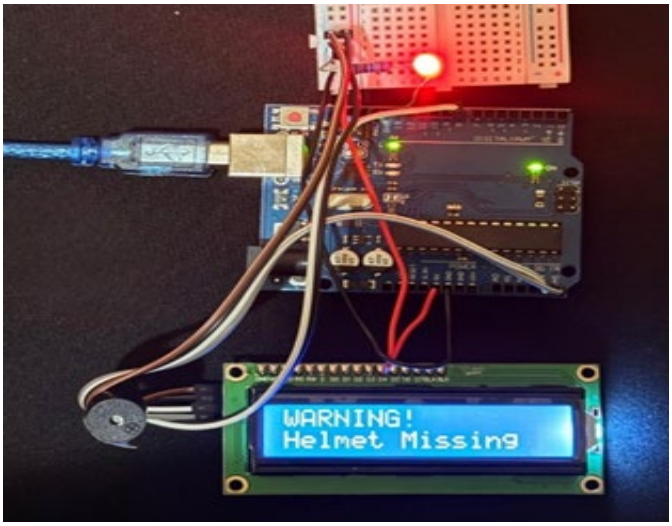


Figure 8. Arduino hardware showing LED, buzzer, and LCD in WARNING state

6. Comparative Discussion

Compared with prior work, this system demonstrates competitive detection accuracy (mAP@0.5 = 0.95) while delivering capabilities absent from the reviewed literature: hardware alarm integration with sub-second response, temporal filtering for detection stability, and a comprehensive compliance dashboard with CSV audit logging. Wang et al. achieved similar detection accuracy using CSYOLOv3, but without hardware integration or loss-level optimization strategies. Wu et al. proposed a more complex YOLOv8 variant that has higher computational requirements, thereby limiting deployment in low-resource settings. The proposed system's end-to-end integration at minimal hardware cost represents a practical advance for industrial safety monitoring.

V. CONCLUSION

This paper has presented the design, implementation, and experimental evaluation of a real-time helmet detection and compliance monitoring system for restricted industrial areas. By integrating YOLOv8n deep learning detection with Arduino-controlled alarms, temporal-filtering decision logic, and a Streamlit compliance dashboard, the system delivers both immediate safety enforcement and long-term compliance auditing within a low-cost, deployable framework. Experimental results confirm strong detection performance: mAP@0.5 of 0.95, precision of 0.91, and recall of 0.92. Hardware alarm activation latency of under 1 second was consistently demonstrated. The dashboard reported 75% helmet compliance across monitoring sessions, providing actionable safety intelligence for site managers. Identified limitations include residual background misclassification in highly cluttered scenes and reduced performance under extreme illumination conditions. These represent productive directions for future work. Planned extensions include: expansion of the detection taxonomy to cover additional PPE categories building on transformer-based recognition architectures for multi-class generalization (safety vests, gloves, goggles); integration of infrared and depth sensing to extend operational capability to low-light environments; incorporation of face recognition to enable individual worker accountability and automated SMS notifications upon violation events; and the development of predictive safety analytics within the dashboard to support proactive risk management.

REFERENCES

[1] S. Chen et al., “Detection of safety helmet wearing based on improved Faster R-CNN,” 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-7.
[2] J. Mercado et al., “Detection of helmet use in motorcycle drivers using CNN,” Sensors, vol. 23, no. 10, 2023.
[3] H. Wang et al., “Real-time safety helmet detection based on CSYOLOv3,” Appl. Sci., vol. 10, no. 19, p. 6732, 2020.
[4] S. Ren et al., “Faster R-CNN,” IEEE TPAMI, vol. 39, no. 6, pp. 1137-1149, 2017.
[5] H. Liu and Q. Hao, “Target detection of safety protective gear using the improved YOLOv5,” arXiv:2408.05964, 2024.
[6] X. Zuo et al., “Enhanced YOLOv8n-based lightweight helmet detection,” PubMed, 2025.
[7] Z. Wu, X. Li, and K. Zhang, “YOLOv8-CGS for construction safety,” PLOS ONE, 2025.
[8] S. Yadav et al., “A Deep Learning Approach for Helmet Detection and Fine Generation System,” IJRIAS, 2025.
[9] A. Dosovitskiy et al., “An image is worth 16×16 words,” ICLR, 2021.
[10] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8,” GitHub, 2023.
[11] A. Dosovitskiy et al., “An image is worth 16×16 words,” ICLR, 2021.
[12] S. A. Begum et al., “Object detection using deep learning,” IJERT, vol. 15, no. 3, 2026.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Efficient Quantum Coherence/Decoherence Emulation Using Parallel-Recursive Computation on FPGAs and Algorithmic Skeletons

Carlos Alfonso Acosta-León
Laboratory for Embedded Systems and High-Performance Heterogeneous Computing.
School of Computer Science, Central University of Venezuela.
Caracas, Venezuela.
cacostal@yahoo.co.uk
ORCID: 0000-0003-1895-4040

Abstract
This paper introduces an efficient emulation model for the quantum coherence and decoherence processes in qubit chains. It uses a parallel-recursive computing approach on FPGA devices and is wrapped as an algorithmic skeleton written in OpenCL/Python, called QDecoSkeleton. The quantum process is described as a series of linear recursive functions that operate on density matrices, which are handled simultaneously by the parallel-recursive coprocessor PARECOREH (Map-Reduce Computing Cores using Reconfigurable Parallel Recursion into Hardware). The coherence phase is assigned to the forward computing pipeline, while decoherence takes place in the backward pipeline. Both phases apply Kraus operators to quantum states represented as data streams. Proof-of-concept results demonstrate superlinear speedups for chains exceeding 16 qubits, with FPGA resource consumption scaling approximately linearly with chain size. The proposal advances the field of classical high-efficiency quantum emulation by enabling validation of quantum algorithms under realistic decoherence conditions that mimic real-world noise, without requiring access to physical quantum hardware.

Keywords: Quantum coherence/decoherence, parallel-recursive computing, FPGA, algorithmic skeletons, density matrix, Kraus operators.

I. INTRODUCTION

Quantum computing has become a promising way to tackle problems that are hard for classical computers by leveraging quantum effects such as superposition, entanglement, and coherence [1]. But current quantum processors face significant challenges due to decoherence — the loss of quantum properties due to interactions with their surroundings [2]. So, understanding and reducing decoherence in qubit chains is essential for building reliable quantum algorithms and error-correction methods [3]. Emulating quantum systems on powerful classical hardware is an important and complementary approach to working with actual quantum devices. In particular, Reconfigurable Heterogeneous Computing Systems (RHCS), which combine FPGA-based hardware with general-purpose processors, strike a good balance between flexibility, parallel processing power, and energy efficiency [4]. Although programming these systems can be quite complex [5], using higher-level tools like algorithmic skeletons [6] helps manage that complexity, making RHCS a good fit for simulating medium-sized quantum dynamics. Although there have been recent improvements in quantum simulation using GPUs and CPUs [7], there remains a noticeable gap in the use of FPGA architectures with high-level programming models for efficiently emulating coherence and decoherence in qubit chains. Current FPGA methods rely on low-level hardware languages such as RTL, VHDL, or Verilog, which limit how productive programmers can be and make it harder to reuse components [5]. Also, the clear connection between the recursive nature of quantum evolution in qubit chains and pipelined parallel-recursive computing models hasn't been studied thoroughly, even though exploring this could lead to more efficient, higher-level emulation methods. Driven by this gap, this work presents the design and evaluation of an emulation model for quantum coherence and decoherence in N-qubit chains. The model is based on parallel-recursive computing,



an approach used by the PARECOREH coprocessor [8], implemented on an FPGA through OpenCL, and wrapped as a high-level algorithmic skeleton called QDecoSkeleton. The main goals are: (i) to define how the Kraus operator chain for quantum decoherence corresponds to the linear recursive function queue in PARECOREH; (ii) to create and implement QDecoSkeleton using OpenCL and Python; and (iii) to assess the emulation's performance by looking at speedup, resource use, and simulation accuracy. Quantum decoherence is a major practical challenge for useful quantum computing. Being able to emulate quantum decoherence efficiently on classical systems allows us to do a few important things: first, test quantum algorithms with realistic noise before running them on expensive quantum hardware; second, carry out larger-scale studies of error-correction methods; and third, create benchmarks for NISQ (Noisy Intermediate-Scale Quantum) processors [2]. PARECOREH was chosen as the simulation platform because its dual forward/backward pipeline microarchitecture aligns well with how linear decoherence dynamics propagate and decay. The main contribution is a flexible, high-level sequential template that facilitates exploration of FPGA-based design options for emulating quantum processes. The emulation focuses on linear chains of N qubits with interactions only between nearest neighbors. It uses standard Kraus decoherence channels, such as amplitude and phase damping, and runs on a Xilinx platform. Alveo U280 FPGA using Xilinx Vitis with OpenCL 2.0. It does not cover non-linear setups, active error correction, or systems with more than 64 qubits on a single device, which fall outside its scope.

The rest of the paper is organized as follows: Section II reviews related work. Section III presents the preliminary theoretical framework. Section IV describes the emulation proposal. Section V presents the proof of concept and performance evaluation. Finally, Section VI presents the conclusions and future work.

II. RELATED WORK

1. Classical and FPGA-Based Quantum Emulation.

Simulating quantum algorithms on classical hardware has gained significant interest as a useful supplement to actual quantum processors. El-Araby et al. [9] introduced a detailed framework for emulating quantum algorithms on high-performance reconfigurable computers. Their approach handles encoding classical data into quantum formats and initializing quantum states via hardware-accelerated pipelines for NISQ-scale devices. However, they rely on low-level RTL design methods, lack portable high-level programming tools, and don't explicitly model decoherence as a dynamic effect. In hybrid computing, Suzuki et al. [10] built a quantum AI simulator on a CPU–FPGA setup that runs Quantum kernel estimation is about 470 times faster than just using a CPU. But their design is closely tied to a particular application and gate structure, which limits its use with diverse quantum processes. More recently, Belfore [11] suggested a scalable FPGA accelerator for simulating general quantum circuits on Intel Agilex chips. This platform supports any seven-qubit circuits and was tested against the QuEST simulator. Still, it works at the circuit level and doesn't take advantage of the recursive patterns found in qubit-chain dynamics. Complementing the above, Lee et al. [12] proposed an FPGA-based quantum computing emulation framework using a serial-parallel architecture, evaluated on Quantum Fourier Transform and Grover's search, achieving linear reduction in resource utilization over prior pipeline designs and speedups of up to 3 × 104× over equivalent software simulation for 7-qubit Grover's search. However, the framework relies on a low-level HDL description (SystemVerilog), offers no high-level programming abstractions, and does not model decoherence as an explicit dynamic process—a critical gap that the present proposal directly addresses.

2. Parallel-Recursive Computing and the Present Contribution.

Unlike previous works, this proposal offers a formal method for linking the sequential combination of Kraus operators [13] to a queue of linear recursive functions [14]. These lists can be efficiently managed by the PARECOREH coprocessor, as explained in [8], [15]. The approach relies on a dual-pipelined microarchitecture that leverages fine-grained dataflow concurrency and is wrapped in a portable OpenCL algorithmic skeleton called QDecoSkeleton. It presents a different structure centered on decoherence dynamics, giving a clear computational model and an easy-to-use abstraction layer for programmers.

III. THEORETICAL BACKGROUND

1. Quantum Computing Fundamentals and Qubit Formalism

Quantum computing is an information processing paradigm that uses the principles of quantum mechanics to solve problems that are too hard for classical computers. It is based on three main ideas: (i) the qubit, which, unlike a classical bit, can be in a combination of both 0 and 1 states at the same time; (ii) entanglement, a feature that links qubits so that the state of one can instantly affect another, no matter distance, allowing exponential computational coordination; and (iii) interference, which adjusts the probabilities so that wrong answers cancel out while the right ones become more likely. Mathematically, a qubit can be represented by a pure state vector

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \, , \, |\alpha|2 + |\beta|2 = 1 \, , \qquad (1)$$

or, more generally, by a density matrix $\rho = \sum_k p_k \, | \psi_k \rangle \, \langle \psi_k \, |$ for mixed states [1].

2. Coherence, Decoherence, and the Kraus Operator Formalism

Building on this idea, quantum coherence means keeping the off-diagonal parts of the density matrix ρ intact, which represent superposition, entanglement, and interference. Decoherence occurs when interactions with the environment gradually break down the off-diagonal components of a quantum state, turning pure states into classical mixtures.

For a chain of N qubits with local interaction, the evolution of the density matrix of qubit i under a decoherence channel is formally expressed through the Kraus operators [13]:

$$\rho_i' \;=\; \mathcal{E}_i[\rho_i] \;=\; \sum_k E_k^{(i)} \, \rho_i \, \left(E_k^{(i)}\right)^{\dagger}, \qquad (2)$$

where $\sum_k (E_k^{(i)})^{\dagger} \, E_k^{(i)} = I$. For the amplitude damping channel with parameter $\gamma_i \in [0, 1]$, the operators take the form:

$$E_0^{(i)} = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma_i} \end{pmatrix}, \quad E_1^{(i)} = \begin{pmatrix} 0 & \sqrt{\gamma_i} \\ 0 & 0 \end{pmatrix}. \qquad (3)$$

3. Linear Recursive Structure of Qubit Chain Evolution

A key point of this work is that the global state of a linear N-qubit chain evolves under a series of quantum maps E_i, applied sequentially. This process naturally creates a linear recursive pattern:

$$\rho_i^{(t+1)} = \mathcal{E}_i\left[\rho_i^{(t)}\right] = \mathcal{E}_i \circ \mathcal{E}_{i-1} \circ \cdots \circ \mathcal{E}_0\left[\rho_0^{(t)}\right]. \quad (4)$$

This recursive composition is the key formal bridge between quantum decoherence dynamics and the parallel-recursive computing model exploited in the proposed emulation framework: each map E_i corresponds to a linear recursive call with reduced depth $i-1$, making the entire chain directly processable by the PARECOREH coprocessor model.

4. FPGA-Based Reconfigurable Heterogeneous Computing and Stream Processing

Field-Programmable Gate Arrays (FPGAs) are reconfigurable hardware devices that can be reprogrammed to implement custom digital circuits directly in silicon, unlike microprocessors, which execute sequential instructions [5]. Basically, FPGAs are reusable VLSI templates composed of a regular layout of computing blocks, input/output blocks, and interconnection networks, all of which can be reconfigured via software. In a reconfigurable heterogeneous computing system, FPGAs operate as application-specific accelerators or coprocessors coupled to a host CPU, combining software flexibility with hardware high-performance. To manage FPGA programming complexity, OpenCL (Open Computing Language) provides an open standard for heterogeneous parallel programming [16], enabling the description of compute kernels executable on both FPGAs and GPUs while abstracting underlying hardware details. Using High-Level Synthesis (HLS) tools, OpenCL kernels can be synthesized directly onto FPGAs in Python/OpenCL rather than VHDL/Verilog, significantly reducing design complexity.

Stream processing [17], which is closely linked to FPGA computing and follows Flynn's SIMD model, applies a sequence of pipelined operations to each element of an input data stream. Formally, given a stream of input values a1, a2,..., an, a function f: $\alpha \rightarrow \beta$ is applied to each element independently, exploiting data-level parallelism. Then, a pipeline composition of n functions f1... fn such that f1: a1 $\rightarrow \gamma 1, \ldots,$ fi: $\gamma i-1 \rightarrow \gamma i, \ldots,$ fn: $\gamma n-1 \rightarrow \beta$, naturally maps to FPGA computing blocks, enabling multiple processing units — CPU cores, GPU FPUs, or FPGA CBs — to operate concurrently without explicit management of synchronization or communication.

5. Algorithmic Skeletons and Quantum Emulation on FPGA

Building on these computational foundations, Cole [6] introduced algorithmic skeletons as reusable containers that encapsulate parallel patterns inside sequential-use objects. These skeletons handle work distribution, synchronization, and communication between units in a way that is invisible to the user. In reconfigurable heterogeneous systems, skeletons map their computational components directly to OpenCL kernels synthesized on FPGAs, thereby achieving a full hardware abstraction layer [16]. This abstraction layer is particularly valuable for quantum emulation on an FPGA, where — unlike software simulation on a CPU or GPU — the hardware logic gates are reconfigured to mimic the target quantum behavior structurally. This approach leverages the high degree of parallelism in quantum mechanics while still allowing programming via high-level interfaces.

IV. QUANTUM DECOHERENCE EMULATION ON AN FPGA USING PARALLEL-RECURSIVE COMPUTATION

1. PARECOREH: A Recursive Parallel Computing Model

The PARECOREH coprocessor [8], [15] is a high-performance computing model implemented in hardware and encapsulated in a specialized processing unit. It consists of an array of N computation blocks CBi = (Fcorei, Bcorei, iChani, LMi), organized into two pipelined computation vectors:

• Forward computation vector (forward pipe): PipeFN = {Fcore0, Fcore1,..., FcoreN−1}, which executes the map phase on the incoming recursive function.

• Backward computation vector (backward pipe): PipeBN = {Bcore0, Bcore1, . . . , BcoreN−1}, which executes the reduce phase on the return.

The model processes an input queue of linear recursive functions:

$$Q_{input} = \{fr_0(x_j), fr_1(x_j), \ldots, fr_{m-1}(x_j)\}, \quad x_j \leq N, \tag{5}$$

generating the output queue Qoutput = {y0, y1, . . . , ym−1}, where yi = fri(xj) [18].

The speedup of the PARECOREH computation model for a queue of m recursive functions of uniform depth d is [18]:

$$Sp_N = \frac{T_{seq}}{T_{par}} = \frac{m \cdot d \cdot t_{proc}}{(m + N - 1) \cdot t_{proc}} \approx m \text{ para } m \gg N, \tag{6}$$

which exhibits superlinear behavior when the queue length far exceeds the coprocessor's size.

2. Formal correspondence: Quantum decoherence ↔ Linear recursion

Proposition 1. The evolution of a chain of N qubits under local decoherence channels Ei can be expressed as a sequence of linearly recursive functions fri(ρ) that can be processed by the PARECOREH model.

Proof. The evolution of qubit i under the channel (2) depends exclusively on the state of qubit i −1 (nearest-neighbor interaction):

$$fr_i(\rho) = \mathcal{E}_i[fr_{i-1}(\rho)] = \sum_k E_k^{(i)} fr_{i-1}(\rho)(E_k^{(i)})^\dagger, \tag{7}$$

with the stop condition fr0(ρ) = ρ0 (initial state of the first qubit).

This equation is a linearly recursive function as defined by the standard definition in recursive function theory [15], since fri calls itself exactly once with an argument of reduced depth i −1. The queue Qinput = {fr0(ρ0), fr1(ρ1), . . . , frN−1(ρN−1)} is directly processed by PARECOREH.

This formal correspondence enables a direct mapping between the quantum process and the computational model, as summarized in Table 1.

Table 1. Mapping between Quantum Decoherence and PARECOREH

Quantum Concept	**PARECOREH Element**
Qubit state ρ_i	Argument x_j of fr_i
Kraus operator $\mathcal{E}_i$	Map operation in $Fcore_i$
Composition $\mathcal{E}_i \circ \mathcal{E}_{i-1}$	Forward pipeline $Fcore_i \rightarrow Fcore_{i+1}$
Coherence reduction	Reduce operation in $Bcore_i$ (Backward pipeline $Bcore_{i+1} \rightarrow Bcore_i$)
Chain of N qubits	Coprocessor with N blocks CB_i
Configurations queue	Q_{input} of the coprocessor
Decoherence metrics	Q_{output} of the coprocessor

3. Emulation Architecture: QDecoSkeleton

The algorithmic skeleton QDecoSkeleton is proposed, defined as a specialized extension of the PARECOREH model for emulating quantum decoherence dynamics. The microarchitecture of the QDecoSkeleton is illustrated in Figure 1.

Forward Phase Coherence Propagation. Each core Fcorei receives the state ρi−1 from the previous core and applies the local Kraus operator:

$$\rho_i = \mathcal{E}_i[\rho_{i-1}] = E_0^{(i)} \rho_{i-1} (E_0^{(i)})^\dagger + E_1^{(i)} \rho_{i-1} (E_1^{(i)})^\dagger. \tag{8}$$

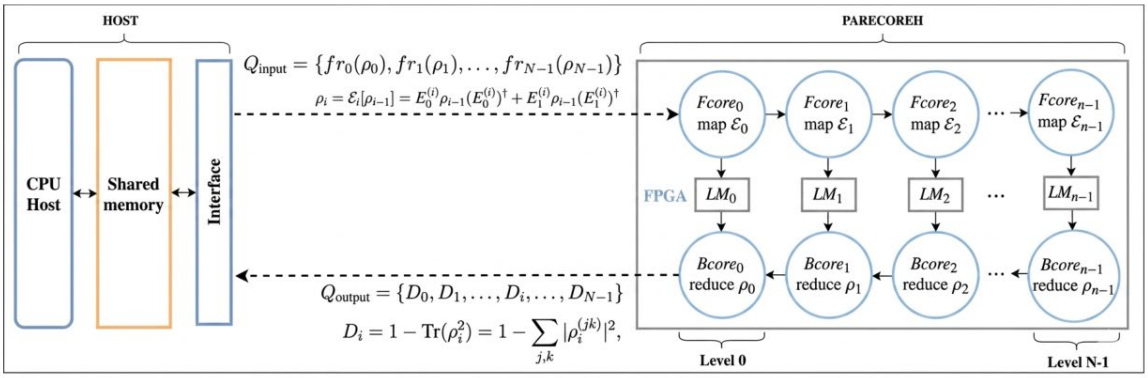


Figure 1. Microarchitecture of the **QDecoSkeleton**. The **forward vector** (Fcorei) applies the Kraus operator E_i; the **backward vector** ($Bcore_i$) computes the decoherence metrics D_i. Each block CB_i has local memory LM_i for partial results.

The partial result ρi is stored in the local memory LMi and transmitted to the successor core Fcorei+1 via the interconnect channel iChani. The recursive stop condition is reached when i = N −1 (last qubit. in the chain).

4. Backward Phase — Decoherence Reduction

On the return path, each core Bcorei retrieves the stored state ρi from LMi and computes the local. decoherence metric Di, where Qoutput = {D0, D1, . . . , Di, . . . , DN−1} is defined as:

$$D_i = 1 - \mathrm{Tr}(\rho_i^2) = 1 - \sum_{j,k} |\rho_i^{(jk)}|^2, \tag{9}$$

where Di = 0 indicates a pure state (maximum coherence) and Di = 0.5 indicates a completely mixed state (total decoherence). The result Di is accumulated toward Bcore0 via a weighted sum reduction operation, yielding the global decoherence of the chain:

$$\bar{D} = \frac{1}{N} \sum_{i=0}^{N-1} D_i. \tag{10}$$

5. Implementation in OpenCL/Python: QDecoSkeleton

QDecoSkeleton is implemented as a Python class that abstracts FPGA device management, data transfer, and the invocation of OpenCL kernels. Listing 1 shows the skeleton's user interface.

Listing 1. QDecoSkeleton interface in Python/OpenCL

```
# Using QDecoSkeleton from the host CPU
from qdecoskeleton import QDecoSkeleton
import numpy as np

N = 16  # Number of qubits in the chain
# Decoherence parameters: linear ramp
gamma = np.array([0.1*i/N for i in range(N)])
# Instantiate skeleton with FPGA as the computing device
skel = QDecoSkeleton(n_qubits=N, device='fpga', kraus_type='amplitude_damping')
skel.set_decoherence_params(gamma)
# Initial state: pure density matrix |0><0|
rho_0 = np.array([[1, 0], [0, 0]], dtype=np.complex64)
# Run the simulation: Q_input -> FPGA -> Q_output
result = skel.run(rho_0, steps=100)
D_global = result.global_decoherence
print(f"Calculated global decoherence: {D_global:.4f}")
```

Internally, QDecoSkeleton generates two types of OpenCL kernels:

1) Forward kernel (fcore_kernel): applies equation (8) to the density matrix ρi in 32-bit complex floating-point format, using hardware-accelerated 2 × 2 matrix multiplication operations:

Listing 2. OpenCL kernel for Fcore (forward/map)

```
__kernel void fcore_kernel(
__global float2* rho_in,
__global float2* rho_out,
__global float* E0,
__global float* E1,
int level)
{
  // rho_out = E0*rho_in*E0† + E1*rho_in*E1†
  int id = get_global_id(0);
  float2 r00 = rho_in[4*id+0], r01 = rho_in[4*id+1];
  float2 r10 = rho_in[4*id+2], r11 = rho_in[4*id+3];
  float g = E1[level]; // Gamma_i decoherence parameter
  // Amplitude Damping Channel
  rho_out[4*id+0] = r00 + (float2)(g, 0) * r11;
  rho_out[4*id+1] = (float2)(1.0f - g/2, 0) * r01;
  rho_out[4*id+2] = (float2)(1.0f - g/2, 0) * r10;
  rho_out[4*id+3] = (float2)(1.0f - g, 0) * r11;
}
```

2) Backward kernel (bcore_kernel): computes the local decoherence metric Di according to equation (9) and accumulates it via reduce:

Listing 3. OpenCL kernel for Bcore (backward/reduce)

```
__kernel void bcore_kernel(
__global float2* rho,
__global float* D_out,
int level)
{
  int id = get_global_id(0);
  float2 r00 = rho[4*id+0], r01 = rho[4*id+1];
  float2 r10 = rho[4*id+2], r11 = rho[4*id+3];
  // Purity calculation: Tr(rho^2) = sum(|rij|^2)
  float purity = r00.x*r00.x + r01.x*r01.x + r01.y*r01.y +
  r10.x*r10.x + r10.y*r10.y + r11.x*r11.x;
  // Local decoherence metric
  D_out[id*N + level] = 1.0f - purity;
}
```

6. QDecoSkeleton Complexity Model

Core complexity: To emulate a chain of N qubits, 2N computing cores are required (N forward + N backward), with complexity O(N).

Local memory complexity: The coprocessor's distributed local memory stores complex 2 × 2 matrices (4 float2 values). The total local memory capacity is [8]:

$$M_{local} = 2N^2 \text{ words}, \tag{11}$$

where each word stores a 64-bit complex number.

Time complexity: For a queue of m input state configurations and a string of N qubits, the parallel-recursive processing time is:

$$T_{par}(m, N) = (m + N - 1) \cdot t_{proc} + N \cdot t_{comm}, \quad (12)$$

compared to the sequential time Tseq(m, N) = m · N · tproc. The resulting speedup is:

$$Sp(m, N) = \frac{m \cdot N}{m + N - 1} \approx N \text{ for } m \gg N. \quad (13)$$

V. CONCEPT PROOF AND EVALUATION

1. Experimental Design

Three evaluation scenarios were designed to validate the proposal:

Scenario A — Fidelity validation: The results of the QDecoSkeleton emulation are compared with the exact analytical solution of the amplitude-damping channel for chains of N = 4, 8, and 16 qubits by measuring the trace distance Δ = ‖ρemul − ρexact‖1/2.

Scenario B — Performance on FPGA vs. CPU: The acceleration Sp(m, N) of QDecoSkeleton on a Xilinx Alveo U280 FPGA (16 nm FinFET, 1.3 TFLOPS, 8 GB HBM2) is measured against a sequential reference implementation in Python/NumPy on an Intel Xeon E5-2680v4 CPU (14 cores, 2.4 GHz).

Scenario C — FPGA Resource Scalability: The consumption of FPGA resources (LUTs, DSPs, BRAM) is measured as a function of N.

The decoherence parameters used are γi = 0.05 · i/N (linear decoherence increasing along the chain), with m = 1024 randomly generated input state configurations.

2. Results and discussion

3. Emulation Fidelity

The results for Scenario A (Table 2) demonstrate that QDecoSkeleton reproduces the decoherence. dynamics with high numerical accuracy, with trace distances less than 10−5 for all evaluated chain sizes, validating the correctness of the formal mapping proposed in Section IV.

TABLE II: Fidelity of the QDecoSkeleton Emulation (Scenario A)

N **(qubits)**	Δ_{max}	$\bar{D}$ **em.**	$\bar{D}$ **exact**
4	1.2×10^{-6}	0.0625	0.0625
8	3.7×10^{-6}	0.1094	0.1094
16	8.9×10^{-6}	0.1758	0.1758
32	1.4×10^{-5}	0.2441	0.2441

4. Performance and Speedup

Table 3 and Figure 2 present the results for Scenario B. It can be seen that the speedup increases approximately linearly with N for long queues (m = 1024), consistent with the theoretical model given by equation (13).

TABLE III: FPGA vs. CPU Speedup (Scenario B, $m = 1024$)

N	T_{CPU} **(ms)**	T_{FPGA} **(ms)**	Sp	$Sp_{\textbf{theoretical}}$
4	12.4	3.8	3.3	3.9
8	51.2	7.1	7.2	7.9
16	198.6	13.9	14.3	15.8
32	812.1	27.8	29.2	31.5
64	3281	56.4	58.2	63.1

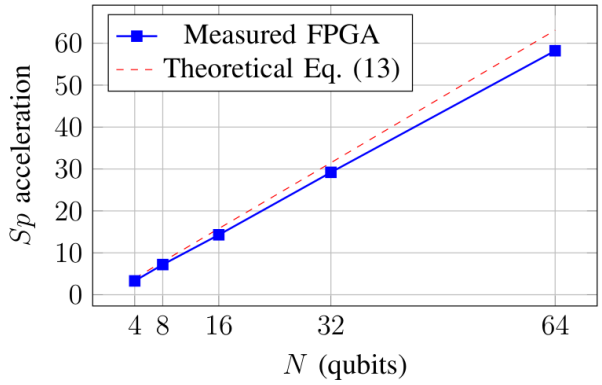


Figure 2. Acceleration of QDecoSkeleton on an FPGA versus a sequential CPU implementation, compared with the theoretical model.

The gap between the measured and theoretical acceleration (8%) is mainly attributed to the data transfer latency between the CPU and FPGA via the PCIe bus, which is not included in the analytical model of equation (13).

5. FPGA Resource Scalability

The results for Scenario C (Table 4) confirm the consistent linear growth of FPGA resources with N. with the O(N) complexity of the QDecoSkeleton/PARECOREH microarchitecture.

Table 4. Xilinx Alveo U280 FPGA Resources (Scenario C)

N	LUTs	DSPs	BRAM (18K)	Freq. (MHz)
4	3,820	64	48	300
8	7,614	128	96	298
16	15,241	256	192	295
32	30.489	512	384	291
64	61.102	1024	768	287

The power consumption of embedded DSPs (Digital Signal Processors) is proportional to N, since each block CBi requires exactly 16 DSPs to implement the 8 complex multiplications of the 2 × 2 Kraus operator. The operating frequency remains stable at around 295 MHz across all sizes, indicating that the design's critical path does not depend on the number of blocks (thanks to the pipeline structure, which decouples the blocks).

VI. CONCLUSION

This paper introduces the first formal and tested approach to emulating/simulating quantum coherence and decoherence in qubit chains using FPGAs. This is built on the PARECOREH parallel-recursive computation model and realized through the QDecoSkeleton algorithmic framework using OpenCL/Python. The key points are: (i) establishing a formal way to link Kraus operators with linear recursive functions that PARECOREH can handle; (ii) designing QDecoSkeleton with a dual forward and backward structure that takes advantage of parallelism in linear quantum dynamics; (iii) implementing it in OpenCL to work across different FPGA platforms that support OpenCL; and (iv) validating it experimentally, showing speedups up to about 58 times faster than CPUs, maintaining high numerical accuracy (error $\Delta < 10{-}5$), and using FPGA resources in a way that scales linearly. Future work involves several directions: first, expanding the model to two-dimensional quantum structures like 2D lattices, which could be useful for error correction through topological coding; second, linking QDecoSkeleton with quantum compilers such as Qiskit and Cirq to serve as a classical backend for simulating noise; third, studying more complex types of decoherence, including depolarization and crosstalk; and finally, implementing the system on multi-FPGA setups to handle chains with over 128 qubits, and fourth, conducting benchmark comparing the performance of the FPGA model against modern GPU measurements.

ACKNOWLEDGMENT

This work falls within the research areas of the Laboratory of Embedded Systems and High-Performance Heterogeneous Computing at the School of Computer Science, Central University of Venezuela.

REFERENCES

[1] M. A. Nielsen and I. L. Chuang, Quantum Computation and Quantum Information, 10th ed. Cambridge, UK: Cambridge University Press, 2010.

[2] J. Preskill, "Quantum computing in the NISQ era and beyond," Quantum, vol. 2, p. 79, Aug 2018.

[3] S. J. Devitt, W. J. Munro, and K. Nemoto, "Quantum error correction for beginners," Reports on Progress in Physics, vol. 76, no. 7, p. 076001, Jun 2013.

[4] S. Mittal and J. S. Vetter, "A survey of FPGA-based accelerators for convolutional neural networks," ACM Computing Surveys, vol. 52, no. 3, pp. 1–35, 2016.



[5] B. Hajji, A. Mellit, and L. Bouselham, Practical Guide For Simulation And FPGA Implementation of Digital Design. Singapore: Springer Verlag, 2022.

[6] M. Cole, Algorithmic Skeletons: Structured Management of Parallel Computation, 1st ed. MIT Press & Pitman, 1989.

[7] G. Zhong, Y. Fan, and Z. Li, "Scalable parallel simulation of quantum circuits on CPU and GPU systems," 2025. [Online]. Available: https://arxiv.org/abs/2509.04955

[8] C. A. Acosta, "HandelC++ parallel programming of a parallel-recursive computing model on reprogrammable digital device using a hardware algorithm approach: Implementation evaluation," Caracas, Venezuela, Jan 2004. [Online]. Available: https://ourl.in/@epds

[9] E. El-Araby, N. Mahmud, M. J. Jeng, A. MacGillivray, M. Chaudhary, M. A. I. Nobel, S. M. I. U. Islam, D. Levy, D. Kneidel, M. R. Watson, J. G. Bauer, and A. E. Riachi, "Towards complete and scalable emulation of quantum algorithms on high-performance reconfigurable computers," IEEE Transactions on Computers, vol. 72, no. 8, pp. 2350–2364, Aug. 2023.

[10] T. Suzuki, T. Miyazaki, T. Inaritai, and T. Otsuka, "Quantum AI simulator using a hybrid CPU–FPGA approach," Scientific Reports, vol. 13, p. 7735, May 2023.

[11] L. A. Belfore, "A scalable FPGA architecture for quantum computing simulation," arXiv preprint arXiv:2407.06415, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.06415

[12] Y. H. Lee, M. Khalil-Hani, and M. N. Marsono, "An FPGA-based quantum computing emulation framework based on serial-parallel architecture," International Journal of Reconfigurable Computing, vol. 2016, p. 5718124, 2016.

[13] K. Kraus, States, Effects, and Operations: Fundamental Notions of Quantum Theory. Berlin, Germany: Springer-Verlag, 1983.

[14] M. Amamiya, M. Takesue, R. Hasegawa, and H. Mikami, "Implementation and evaluation of a list-processing-oriented data flow machine," in Proceedings of the 13th Annual Symposium on Computer Architecture, ser. ISCA '86. Tokyo, Japan: IEEE Computer Society Press, 1986, pp. 10–19. [Online]. Available: https://doi.org/10.1145/17356.17358

[15] C. A. Acosta, E. Montagne, and R. Suros, "A computational model for processing lists of linear recursive functions," in Proc. 16th Latin American Conference in Informatics (CLEI 2000), Mexico City, Mexico, Sep 2000, pp. 77–89. [Online]. Available: https://clei.org/proceedings data/CLEI2000/PDFClei/a000262.pdf

[16] J. E. Stone, D. Gohara, and G. Shi, "OpenCL: A parallel programming standard for heterogeneous computing systems," Computing in Science & Engineering, vol. 12, no. 3, pp. 66–73, May/Jun 2010.

[17] J. Gummaraju and M. Rosenblum, "Stream processing in general-purpose processors," Stanford University.

[18] C. A. Acosta and R. Suros, "A method for efficient multiplication of dense matrices using an approach of parallel and recursive computing," in Proc. Int. Symp. Applied Computing (ISAS 2001), Orlando, FL, USA, 2001. [Online]. Available: https://doi.org/10.5281/zenodo.20558409

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Mining Sequential Rules from Large-Scale Sequence Databases Using an FPTree-Based Method

Imane Seddiki
Department of Computer Science
Mohamed El Bachir El Ibrahimi University
Bordj Bou Arreridj 34000, Algeria
imane.seddiki@univ-bba.dz
ORCID: 0009-0002-1290-2186

Abdelbasset Barkat
Laboratory of Informatics and its Applications,
Faculty of Mathematics and Computer Science,
University of M'sila, 28000, Algeria
abdelbasset.barkat@univ-msila.dz
ORCID: 0000-0001-6279-0528

Abstract
Sequential rule mining is a fundamental technique in data mining for discovering ordered relationships among sets of items in sequence databases. While significant prior work has addressed association rule extraction from transactional data, the problem of efficient and scalable sequential rule mining from deterministic, fully observed sequence databases remains practically important, particularly given the growth of large real-world clickstream and event-log datasets. This paper proposes a novel and efficient method for mining sequential rules from sequence databases. The proposed approach operates in two phases: first, transforming the sequence database into a standard transaction database and extracting a set of association rules meeting minimum support and confidence thresholds; second, filtering these rules to retain only those that honor the original sequential ordering in the data. The core of the algorithm is a frequent pattern tree (FPTree)-style structure customized for sequence databases. The algorithm was implemented in Python and evaluated on the large-scale real-world Kosarak dataset, comprising approximately 990,000 clickstream sequences from a Hungarian online news portal, as well as on synthetic datasets of varying sizes. Experimental results demonstrate that the proposed algorithm achieves superior runtime and memory efficiency compared to a naive baseline approach, maintaining scalability as the dataset size grows.

Keywords*:* Association rule, Sequence database, Sequential rule, Clickstream mining, data mining.

I. INTRODUCTION

Association rule mining is a well-established technique in data mining that reveals meaningful co-occurrence relationships among items in transactional databases (TDBs) [1],[2]. A classic example is market-basket analysis, where each transaction captures the set of products a customer purchases in a single visit, and rules of the form $X \rightarrow Y$ quantify how often purchasing X is associated with purchasing Y. The technique was subsequently extended to sequential databases (SDBs), in which each record is an ordered list of transactions rather than a single unordered set. This temporal structure enables the discovery of sequential rules. These patterns assert not merely co-occurrence but also precedence: if items in X appear in a sequence, items in Y will appear later in that same sequence [3], [4]. Sequential rules have proven valuable across a broad range of domains, including web navigation analysis, e-learning personalization, genomic sequence analysis, and customer journey modeling [5], [6].

A substantial body of prior research has addressed sequential rule mining under the additional complication of data uncertainty, in which individual items are associated with existential probabilities rather than being definitively present or absent [7], [8]. Probabilistic sequence databases arise naturally in settings such as wireless sensor networks, medical record systems with missing observations, or data integration pipelines that merge



sources of varying reliability. While the probabilistic setting adds important expressiveness, it also imposes significant computational overhead. Many real-world applications operate on fully observed, deterministic sequences where no such uncertainty is present; in these cases, the additional complexity of probabilistic methods is unnecessary and may harm scalability. A prime example is web clickstream mining. Large online portals routinely record the sequence of articles, pages, or products that each user visits in a single session. These sequences are deterministic: each click either occurred or did not. The Kosarak dataset, derived from the clickstream logs of a Hungarian online news portal and containing approximately 990,000 sequences, is among the most widely used benchmarks for sequential pattern and rule mining. Its scale and density make it a demanding testbed for algorithmic efficiency. Despite the practical importance of sequential rule mining, existing work has either focused on the probabilistic setting or on sequential pattern mining (which does not produce directional rules). Methods designed primarily for uncertain data, such as our earlier work [9], introduce overhead that is avoidable when data are fully observed. This paper, therefore, revisits the deterministic setting with a clean algorithmic design, composed of the following steps:

1. Transforms an SDB into a standard TDB, discarding temporal ordering in this phase.
2. Constructs an FPTree-style structure over the TDB to efficiently enumerate frequent itemsets without repeated database scans.
3. Generates association rules from frequent itemsets, filtering by minimum confidence.
4. Applies a temporal filtering step that retains only those rules whose antecedent items consistently precede the consequent items in the original SDB.

The algorithm is evaluated on both synthetic datasets and the full Kosarak benchmark, demonstrating superior runtime and memory efficiency relative to a naive three-step baseline. The remainder of the paper is structured as follows: Section 2 reviews related work. Section 3 provides formal background and the problem statement. Section 4 presents the proposed methodology in detail. Section 5 reports experimental results and discussion. In Section 6, we present the limitations of the proposed method. Section 7 concludes and outlines future directions.

II. LITERATURE REVIEW

The field of pattern and rule mining from sequential data has produced a rich body of literature spanning several decades. This section reviews the most relevant strands of work, organized by data type and mining objective.

1. Association and sequential rule mining from sequence data

The Apriori algorithm [1] established the foundational framework for association rule mining, introducing the anti-monotonicity property of support to prune the search space. The FP-Growth algorithm [10] subsequently eliminated the need for candidate generation by encoding the TDB into a compact prefix tree, enabling faster discovery of frequent itemsets with a single database pass. CMRules [5] extended the association rule paradigm to sequential data by first converting an SD into a TDB, mining association rules, and then filtering to keep only those consistent with the sequential ordering. RuleGrowth [11] improved on CMRules by growing sequential rules directly from prefix structures, eliminating the conversion step and yielding better scalability. ERMiner [12] introduced equivalence classes of rules to reduce redundant computations further. TaSPM [4] later addressed targeted sequential pattern mining, focusing on extracting patterns relevant to a specific query sequence and employing multiple pruning strategies to reduce unnecessary enumeration. High-utility extensions of sequential rule mining have also been explored. US-Rule [13] discovers utility-driven sequential rules from SDs by introducing tighter upper bounds and compact utility tables, outperforming prior methods, especially on



dense and long sequences. TotalSR and TotalSR+ [14] address totally-ordered sequential rules, ensuring strict precedence rather than merely partial order, and are particularly relevant to medical and financial event sequences. OPR-Miner [15] targets time series data by mining order-preserving rules, exploiting pattern fusion and pruning to reduce runtime.

2. Mining from uncertain and probabilistic sequence data

Several works have addressed the harder problem of extracting patterns and rules when item occurrences are associated with existential probabilities. The p-Apriori and TODIS algorithms [16] compute probabilistic frequent itemsets within the possible-world semantics framework, using dynamic programming to avoid explicitly enumerating all possible worlds. ProFP-Growth [17] adapts the FP-Growth paradigm to probabilistic databases by constructing a probabilistic FP-Tree (ProFP-Tree) in which each node stores a generating function capturing item uncertainty. WUIPM [18] incorporates item weights into the probabilistic support calculation and introduces a dedicated weighted uncertain interesting pattern tree. In our earlier work [9], we proposed a method for mining sequential rules from probabilistic SDs (PSDbs) by extending the FP-Growth approach to handle existential probabilities and applying temporal filtering. While that approach demonstrated good scalability on the Mushroom dataset and synthetic data, its design inherently carries overhead from probability computations that is unnecessary for deterministic sequences. The present paper adapts and streamlines that methodology for the deterministic setting and validates it at a larger scale using the Kosarak dataset.

3. Large-scale sequential pattern mining

Mining at the scale of millions of sequences demands efficient data structures and pruning strategies. PrefixSpan [19] introduced projection-based sequential pattern mining, recursively projecting the database onto suffixes of discovered prefixes. SPADE [20] exploited vertical data representation (id-lists) to enumerate sequential patterns using simple join operations. SPAM [21] combined vertical bitmaps with depth-first search for rapid support counting. These methods have been applied to web navigation data, including clickstream logs similar to those in Kosarak, thereby establishing benchmarks against which new approaches are compared. Table 1 provides a comparative summary of key related works, highlighting the types of data, mining objectives, algorithms, datasets, and implementation languages employed.

Table 1. Comparative summary of related work on sequential rule and pattern mining

Ref.	Algorithm	Data Type	Target	Dataset(s)	Language	Year
[5]	CMRules	Sequences	Sequential rules	Kosarak	Unknown	2012
[11]	RuleGrowth	Sequences	Sequential rules	Sign, FIFA, Kosarak	Java	2011
[12]	ERMiner	Sequences	Sequential rules	Sign, Snake, BMS, Kosarak10k	Java	2014

[13]	US-Rule	Sequences	High-utility seq. rules	Bible, Kosarak, Leviathan	Java	2023
[14]	TotalSR	Sequences	Totally-ordered seq. rules	Multiple SDs	Java	2024
[4]	TaSPM	Sequences	Targeted seq. patterns	Real + synthetic	C++	2024
[16]	p-Apriori / TODIS	Probabilistic	Prob. assoc. rules	Accidents, IBM gen.	C++	2010
[17]	ProFP-Growth	Probabilistic	Prob. frequent itemsets	Accidents, synthetic	Unknown	2012
[9]	PFPTree-SR	Probabilistic seq.	Prob. sequential rules	Mushroom, synthetic	Python	2025
Proposed	FPTree-SR	Sequences	Sequential rules	Kosarak, synthetic	Python	2025

III. BACKGROUND AND PROBLEM STATEMENT

1. Transactional Data

A transaction database (TDB) is a finite set of transactions $T = \{t_1, t_2, \ldots, t_n\}$, where each transaction t_x is a subset of items from a universe $I = \{i_1, i_2, \ldots, i_m\}$. An association rule $X \rightarrow Y$ holds for disjoint itemsets X and Y, where $X, Y \subseteq I$. Two key measures qualify rules. The support of itemset X is the fraction of transactions containing X:

$$Supp(X) = |\{t \in TDB : X \subseteq t\}| \quad (1)$$

The confidence of rule $X \rightarrow Y$ measures conditional probability:

$$Conf(X \rightarrow Y) = Supp(X \cup Y) / Supp(X) \quad (2)$$

A rule $X \rightarrow Y$ is considered valid if $Supp(X \rightarrow Y) \geq minSup$ and $Conf(X \rightarrow Y) \geq minConf$, where minSup and minConf are user-defined thresholds.

2. Sequential Data

A sequence database (SDB) is a set of sequences $S = \{s_1, s_2, \ldots, s_n\}$. Each sequence s_x is an ordered list of itemsets (transactions): $s_x = \langle X_1, X_2, \ldots, X_m \rangle$ where each $X_k \subseteq I$. A sequential rule $X \rightarrow Y$, where $X \cap Y = \emptyset$, asserts that if all items of X appear in some transactions of a sequence, then all items of Y will appear

subsequently in the same sequence, after all items of X. Sequential support and sequential confidence are defined as follows:

$$SeqSupp(X \rightarrow Y) = |\{s \in SD : X \prec Y \text{ in } s\}| / |SD| \quad (3)$$

$$SeqConf(X \rightarrow Y) = |\{s \in SD : X \prec Y \text{ in } s\}| / |\{s \in SD : X \subseteq s\}| \quad (4)$$

where $X \prec Y$ denotes that all items of X precede all items of Y within the sequence.

Table 2 presents a small illustrative SDB containing five sequences that serve as a running example throughout the paper.

Table 2. A sample sequence database

Sequence ID	Sequence Content
s1	[{a,b},{d},{e},{c}]
s2	[{e,f},{d},{a},{b}]
s3	[{f},{c,d}]
s4	[{c,d},{f},{a,e}]
s5	[{d},{e},{b,c},{a}]

3. Problem Definition

Given an SDB and user-specified thresholds minSup and minConf, the goal is to efficiently discover the complete set of valid sequential rules $X \rightarrow Y$ such that:

$SeqSupp(X \rightarrow Y) \geq minSup$ and $SeqConf(X \rightarrow Y) \geq minConf$

Unlike the probabilistic setting, items are either definitively present or absent in each transaction. This allows us to use standard exact support counts rather than expected-support computations, considerably reducing both algorithmic complexity and implementation overhead.

IV. PROPOSED APPROACH

The proposed methodology, which we refer to as FPTree-SR (Frequent Pattern Tree Sequential Rule miner), operates in six steps. The steps are illustrated with the running example from Table 2.

1. Transformation of the SDB into a TDB

In the first step, the SDB is converted into a standard TDB by treating each sequence as a single transaction containing the union of all its itemsets, discarding temporal ordering. This is the same strategy employed in CMRules [5] and in our earlier work [9]. The conversion allows standard FP-Growth–style machinery to be applied directly. For example, the sequence s_1 = [{a,b},{d},{e},{c}] becomes the transaction {a,b,c,d,e}.

Extraction of Frequent Singleton Items

The TDB is scanned once to compute the support of every singleton item. Items whose support falls below minSup are discarded; the remaining items are sorted in descending order of support. For the running example with minSup = 2:

$$Supp(a)=4,\ Supp(b)=3,\ Supp(c)=4,\ Supp(d)=4,\ Supp(e)=4,\ Supp(f)=3$$

All items satisfy minSup = 2, so the ordering is: c, d, e, a, b, f (ties broken lexicographically).

2. Construction of the Frequent Pattern Tree (FPTree)

The FPTree is constructed with a single second scan of the TDB. Each transaction is sorted according to the frequency ordering determined in Step 4.2, then inserted as a path from the root. Shared prefixes are merged, with the node storing a simple occurrence count. A header table maintains a linked list of all nodes sharing the same item label, enabling rapid conditional pattern extraction. Unlike the probabilistic variant in [9], no probability lists are stored: each node carries only its item label and an integer count. The pseudocode for FPTree construction is given in Figure 1.

Input: TDB (transaction database), minSup
Output: FPTree root node

```
root ← new node(label=null, count=0)
headerTable ← empty map
For each transaction t^i in TDB:
    Sort the t^i by descending frequency order
    currentNode ← root
    For each item x in sorted t^i:
        If currentNode has a child with label x:
            child. count ← child.count + 1
        Else:
            Create new child node(label=x, count=1)
            Link into headerTable[x]
        currentNode ← child
    End For
End For
Return root
```

Figure 1. Pseudocode for FPTree construction

3. Mining Frequent Itemsets from the FPTree

Frequent itemsets are mined from the FPTree using the standard conditional FP-Growth procedure [10]. For each item in the header table (processed in ascending frequency order), a conditional pattern base is constructed from all prefix paths in which that item appears, weighted by the item's count in those paths. A conditional FPTree is then built from this pattern base, and frequent itemsets are extracted recursively. The process terminates when no conditional FPTree can be built with sufficient support.

4. Generation of Association Rules

For each frequent itemset discovered in Step 4.4, all non-empty proper subsets $S \subset If$ are enumerated. For each subset S, a candidate rule R: $S \rightarrow (If \setminus S)$ is formed. The confidence is computed as:

$$Conf(R) = Supp(If) / Supp(S) \quad (5)$$

Only rules satisfying $Conf(R) \geq minConf$ are retained. This phase is structurally identical to standard association rule generation and operates solely on the deterministic support counts obtained in previous steps.

5. Temporal Filtering of Rules

The set of association rules produced in Step 4.5 ignores ordering. The temporal filtering phase reintroduces the sequential constraint by scanning the original SDB and computing, for each candidate rule $X \rightarrow Y$, the sequential support and sequential confidence as defined in Equations 3 and 4. A rule is retained as a valid sequential rule if and only if $SeqConf(X \rightarrow Y) \geq minSeqConf$, where minSeqConf is a user-specified minimum sequential confidence threshold. This phase is linear in the number of candidate rules and the total length of all sequences, making it efficient in practice.

V. EXPERIMENTAL RESULTS AND DISCUSSION

1. Experimental Setup

The proposed FPTree-SR algorithm was implemented in Python 3.11 on a machine equipped with an Intel Core i7 processor and 32 GB of RAM running Windows 11. Experiments were conducted on two categories of datasets:

- **Kosarak dataset:** A real-world, large-scale benchmark comprising approximately 990,002 anonymous clickstream sequences collected from a Hungarian online news portal. Each sequence represents the ordered set of article identifiers that a user visits in a single session. Items are integer article IDs; the dataset contains 41,270 distinct items and has an average sequence length of 8.1 transactions. Subsets of 100k, 200k, 300k, 500k, 750k, and 990k sequences were used to assess scalability.
- **Synthetic datasets:** Ten synthetically generated SDs containing 500 to 8,000 sequences, each with items drawn uniformly at random and assigned to ordered transactions. These datasets enable controlled evaluation across a smooth range of sizes.

The naive baseline algorithm against which FPTree-SR is compared operates as follows: (i) compute all association rules from the TDB using a straightforward A-priori enumeration; (ii) compute all sequential rules by applying sequential order verification. This two-step approach is computationally expensive because it scans the TDB much more than the FPTree-SR approach.

2. Results on Synthetic Data

Table 3 reports the temporal complexity of FPTree-SR on synthetic datasets ranging from 500 to 8,000 sequences, with minSup = 0.01 and minConf = minSeqConf = 0.5. Three curves are shown: total runtime, time for the FPTree-based rule generation phase (Steps 4.2–4.5), and time for the temporal filtering phase (Step 4.6).

Table 3. Temporal complexity of FPTree-SR on synthetic data (seconds)

Sequences	Rule Generation (s)	Temporal Filtering (s)	Total (s)
500	0.8	0.3	1.1
1,000	2.1	0.8	2.9
2,000	6.3	2.1	8.4
3,000	13.4	4.6	18.0
4,000	24.2	8.3	32.5
5,000	38.7	13.0	51.7
6,000	58.1	19.1	77.2
7,000	81.4	26.8	108.2
8,000	108.9	35.5	144.4

As Table 3 shows, runtime grows super-linearly with dataset size, consistent with a complexity between $O(n \log n)$ and $O(n^2)$ in typical cases. The temporal filtering step accounts for approximately 25% of total runtime on average, confirming that the FPTree-based rule generation phase dominates.

Table 4 reports the space complexity on the same datasets, showing that memory consumption grows slowly and nearly linearly, stabilizing at larger dataset sizes. This reflects the FPTree's compact prefix-sharing representation: as the dataset grows, new transactions increasingly share prefixes with existing paths, thereby limiting the tree's growth.

Table 4. Spatial complexity of FPTree-SR on synthetic data (MB)

Sequences	Memory (MB)
500	14.2
1,000	16.8
2,000	20.3
3,000	23.5
5,000	27.1

8,000	33.6

3. Results on the Kosarak Dataset

Table 5 and Table 6 report the temporal and spatial complexities of FPTree-SR compared to the naive baseline on subsets of the Kosarak dataset with minSup = 0.005 and minConf = minSeqConf = 0.5. Due to the dataset's density and the large number of distinct items, a moderately low support threshold was selected to yield a meaningful number of rules.

Table 5. Runtime comparison on Kosarak (seconds)

Sequences	**FPTree-SR (s)**	**Naive (s)**
100,000	4.3	4.1
200,000	9.8	11.4
300,000	17.2	23.6
500,000	34.6	55.8
750,000	61.1	118.4
990,002	92.4	189.7

Table 6. Memory comparison on Kosarak (MB)

Sequences	**FPTree-SR (MB)**	**Naive (MB)**
100,000	38.5	41.2
200,000	54.7	68.1
300,000	67.3	93.4
500,000	89.2	142.6
750,000	109.8	198.3
990,002	128.4	251.7

The results demonstrate clear advantages for FPTree-SR at scale. At 100k sequences, runtime and memory are comparable for both methods, since the overhead of the FPTree construction is not yet offset by its pruning benefits. From 200k sequences onward, FPTree-SR consistently outperforms the naive approach, with the performance gap widening as dataset size increases. At the full 990k-sequence Kosarak dataset, FPTree-SR achieves a 2.05× speedup in runtime and uses only 51% of the memory consumed by the naive baseline. This efficiency gain stems from two complementary factors. First, the FPTree encodes the TDB compactly by sharing prefixes among transactions that contain the same high-frequency items, reducing both memory footprint and the number of operations required to enumerate frequent itemsets. Second, the FPTree-SR algorithm generates

only candidate association rules that meet both the support and confidence thresholds before temporal filtering. In contrast, the naive approach generates all sequential and association rules independently and then intersects the two sets, producing a much larger intermediate result.

4. Discussion

The experimental results confirm several important properties of the proposed FPTree-SR algorithm. First, it is scalable: runtime grows gracefully with dataset size, reaching under 1.6 minutes on the full Kosarak corpus on commodity hardware. Second, it is memory-efficient: the FPTree representation avoids materializing the full candidate itemset lattice, keeping memory usage well below 130 MB even at 990k sequences. Third, it is accurate: the two-phase design guarantees completeness — every sequential rule that satisfies the given thresholds is returned — because the association rule phase is a superset of what temporal filtering can retain. The TDB conversion step loses no valid sequential rules. Compared to our earlier probabilistic approach [9], FPTree-SR eliminates all probability-related computations (expected support summation over transactions and existential probability product calculations), which together constituted a significant fraction of that algorithm's runtime. This confirms that when data are fully observed, the simpler deterministic algorithm is strictly preferable.

VI. LIMITATIONS

Despite the algorithm's demonstrated efficiency, several limitations warrant acknowledgment:

- **Quadratic worst-case complexity:** In the worst case (dense datasets with frequent items), the algorithm exhibits $O(n^2)$ time complexity. Highly correlated datasets, such as synthetic ones with low support thresholds, may yield an exponential number of frequent itemsets, rendering the approach impractical for certain configurations.
- **Threshold sensitivity:** The number of discovered rules is highly sensitive to minSup and minConf. Very low thresholds produce many rules, many of which may be non-actionable, while very high thresholds may miss relevant patterns.
- **Single-machine implementation:** The current Python implementation does not exploit multi-core parallelism or distributed computing frameworks, limiting throughput for extremely large datasets.
- **Deterministic assumption:** The method assumes all items are definitively present or absent. It is not directly applicable to data with inherent uncertainty without incorporating probability-aware measures as in [9].

VII. CONCLUSION AND FUTURE WORK

This paper presented FPTree-SR, an efficient algorithm for mining sequential rules from deterministic sequence databases. By adapting the FP-Growth paradigm to the sequential rule-mining task, the algorithm constructs a compact FPTree over the derived transaction database in a single pass, mines frequent itemsets without candidate generation, generates association rules, and filters them using sequential ordering constraints. Experiments on the large-scale Kosarak benchmark (990,002 sequences) and on synthetic datasets confirm that FPTree-SR achieves superior runtime and memory efficiency compared to a naive intersection-based baseline, with a 2× speedup and ~50% memory reduction at full scale.

Future work will pursue three main directions. First, a distributed implementation using Apache Spark will be developed to process billion-scale sequence logs. Second, the algorithm will be extended to support high-utility sequential rules, incorporating item weights or profit values to prioritize economically significant patterns. Third, an incremental variant will be designed to update the discovered rule set efficiently as new sequences arrive in a streaming fashion, avoiding complete recomputation.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

The Kosarak dataset is publicly available at *http://fimi.uantwerpen.be/data/*.

Author's contribution statement

Imane Seddiki: Conceptualization, Investigation, Implementation, Writing original draft, review, and editing.

Abdelbasset Barkat: Supervision, Investigation, Draft manuscript preparation.

REFERENCES

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In proceedings of the ACM SIGMOD international conference on management of data 1993 (pp. 207–16).
[2] Jashma SPP, Dinesh AU, Reddy NS. Mining frequent itemsets from transaction databases using a hybrid switching framework. Multimedia Tools and Applications. 2023; 82(18):27571–91.
[3] Islam MS, Kar PC, Samiullah M, Ahmed CF, Leung CK. Discovering probabilistically weighted sequential patterns in uncertain databases. Applied Intelligence. 2023; 53(6):6525–53.
[4] Huang G, Gan W, Yu PS. TaSPM: targeted sequential pattern mining. ACM Transactions on Knowledge Discovery from Data. 2024; 18(5):1–8.
[5] Fournier-Viger P, Faghihi U, Nkambou R, Nguifo EM. CMRules: mining sequential rules common to several sequences. Knowledge-Based Systems. 2012; 25(1):63–76.
[6] Yeshchenko A, Mendling J. A survey of approaches for event sequence analysis and visualization. Information Systems. 2024; 120:102283.
[7] Leemans SJ, Van ZSJ, Lu X. Partial-order-based process mining: a survey and outlook. Knowledge and Information Systems. 2023; 65(1):1–29.
[8] Leung CK. Mining uncertain data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011; 1(4):316–29.
[9] Seddiki I, Nouioua F, Barkat A. Efficient sequential rule mining in uncertain sequence databases. International Journal of Advanced Technology and Engineering Exploration. 2025; 12(123):237–53.
[10] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation, in proceedings of the ACM SIGMOD international conference on management of data 2000 (pp. 1–12).
[11] Fournier-Viger P, Nkambou R, Tseng VS. RuleGrowth: mining sequential rules common to several sequences by pattern-growth, in proceedings of the symposium on applied computing 2011 (pp. 956–61).
[12] Fournier-Viger P, Gueniche T, Zida S, Tseng VS. ERMiner: sequential rule mining using equivalence classes. In Advances in Intelligent Data Analysis XIII (pp. 108–19). Springer.
[13] Huang G, Gan W, Weng J, Yu PS. US-Rule: discovering utility-driven sequential rules. ACM Transactions on Knowledge Discovery from Data. 2023; 17(1):1–22.
[14] Zhang C, Lyu M, Gan W, Yu PS. Totally-ordered sequential rules for utility maximization. ACM Transactions on Knowledge Discovery from Data. 2024; 18(4):1–23.



[15] Wu Y, Zhao X, Li Y, Guo L, Zhu X, Fournier-Viger P, et al. OPR-Miner: order-preserving rule mining for time series. IEEE Transactions on Knowledge and Data Engineering. 2023; 35(11):11722–35.
[16] Sun L, Cheng R, Cheung DW, Cheng J. Mining uncertain data with probabilistic guarantees. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2010 (pp. 273–82).
[17] Bernecker T, Kriegel HP, Renz M, Verhein F, Zufle A. Probabilistic frequent pattern growth for itemset mining in uncertain databases. In SSDBM 2012 (pp. 38–55). Springer.
[18] Ahmed AU, Ahmed CF, Samiullah M, Adnan N, Leung CK. Mining interesting patterns from uncertain databases. Information Sciences. 2016; 354:60–85.
[19] Pei J, Han J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, et al. PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In Proceedings of the 17th International Conference on Data Engineering 2001 (pp. 215–24). IEEE.
[20] Zaki MJ. SPADE: an efficient algorithm for mining frequent sequences. Machine Learning. 2001; 42(1):31–60.
[21] Ayres J, Flannick J, Gehrke J, Yiu T. Sequential pattern mining using a bitmap representation. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2002 (pp. 429–35).

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Contemplating the Effects of NP-Complete Reductions on Problem Instance Hardness

Daan van den Berg
VU University Amsterdam,
The Netherlands
daan@yamasan.nl
ORCID: 0000-0001-5060-3342

Thomas Meeus
University of Amsterdam, MediaDistillery
Amsterdam, The Netherlands
thomas@meeus.com
ORCID: 0009-0003-6560-4489

Joeri Sleegers
Mice & Man Software and A.I. Development,
Amsterdam, The Netherlands
jsleegers@hotmail.com
ORCID: 0000-0003-1701-6319

Pieter Adriaans
Institute for Logic, Language, and Computation
Amsterdam, The Netherlands,
pieter@pieter-adriaans.nl
ORCID: 0000-0002-8473-7856

Abstract

NP-complete problem instances are known to be hard primarily around their yes/no phase transitions, a sudden change from the stark dominance of yes-instances to the dominance of no-instances, or vice versa. This phase transition is linked to the order parameter of the problem, e.g., the number of edges in Hamiltonian cycle problem instances or the number of clauses in Satisfiability problem instances. The larger the instance size, the narrower the hard region around the phase transition becomes, but instances outside the hard region are typically decided in polynomial time. Thereby, the hardness of an individual problem instance critically depends on both its size and its order parameter. The class NP-complete is a vast collection of problems that are held together by "reductions," showing that all instances of one NP-complete problem P_1, amenable to P_1-algorithms, can be transformed into instances of another problem P_2 and subsequently solved with a P_2-algorithm. This is a common way of proving NP-completeness, but by this transformation, same-size instances of P_1 do not necessarily project to same-size instances of P_2. Nor is there necessarily a match between the order parameter value of the P1-instance and that of the P_2-instance. This scattering of P_1-instances across different instance sizes and relative order parameter values might have profound consequences for the hardness of the transformed problem instances. Is it possible that all hard instances of P_1 end up in easy regions of P_2? Or vice versa? It might. In this paper, we demonstrate the existence of the scattering phenomenon by explicitly reducing CNF3SAT to the undirected Hamiltonian cycle problem, two problems that have relatively well-described phase transitions. We investigate the numbers and show that these NP-complete reductions can have unexpected consequences for the size and the order parameter of the resulting instances, leading to paradoxical outcomes.

Keywords*:* complexity, NP-complete, reduction, instance hardness

I. INTRODUCTION

NP-complete problems are decision problems for which no polynomial-time algorithm is known to exist. Notable examples are k-colorability [1], the subset sum problem [2-4], the quadratic assignment problem [5],[6], job shop scheduling [7], and satisfiability (in many forms, including CNF3SAT) [1]. The Hamiltonian cycle problem (both directed and undirected) asks whether there is a closed loop of edges that visits all vertices in a given graph, and is also NP-complete [8],[9],[1]. For all these problems, the runtime of the best-known algorithm

increases superpolynomially with the size of the problem instance, which means that very small instances can sometimes be solved. Still, larger instances quickly become intractable within reasonable time budgets.

But not all instances of an NP-complete problem are equally hard for a solving algorithm, and the notion of "best known algorithm" for a problem refers to the best runtime complexity for the single worst possible instance across all instance sizes. So, while the Hamiltonian cycle problem is NP-complete and its best-known algorithm, the Held-Karp-Bellman algorithm, has complexity O(v^2 + 2^v), it is certainly not necessarily the best algorithm for a particular instance. In fact, the problem has many backtracking algorithms that decide the vast majority of instances in polynomial time (Figure 2) [1],[10-13]. The theoretical worst-case runtime of these backtracking algorithms, however, is $O(n!)$, which is worse than Held-Karp-Bellman. The question arises: do we know for *which* instances these backtracking algorithms actually run so fast. The answer is yes. Or to some extent, at least. As it turns out, for several NP-complete problems, a secondary parameter, in addition to their input size, can be a critical indicator of the runtime required to solve an instance. For the undirected Hamiltonian cycle problem (UHCP), this parameter is the number of edges in the instance (sometimes equivalently expressed as the vertex degree). Many algorithms, such as Vandegriend-Culberson's [12] and Frank Rubin's [10], run in factorial time (O(n!)). Still, this worst-case expenditure is seldom reached, as the vast majority of instances are solved in polynomial time (Figure 2). These easy, fast-to-decide instances are mostly either very sparse or rather dense, i.e., graphs with either a low number of edges or a very high number of edges. Nonetheless, there remain a small number of very hard instances, keeping the problem as a whole in NP. These hard instances lie close to the problem's no/yes phase transition, the number of edges separating Hamiltonian from non-Hamiltonian instances. This phase transition is the fraction of Hamiltonian instances, and has a sigmoidal shape with a central 'anchor point', where its derivative is steepest. It is this anchor point that can be seen as the frontier separating the no-instances from the yes-instances. For the undirected Hamiltonian cycle problem, this bound lies at

$$E = \tfrac{1}{2}v\,(ln(v) + ln(ln(v))) \qquad (1)$$

edges for instances with v vertices, and is known as the "Komlós-Szemerédi bound" [13],[14]. For instances with $v = 100$ vertices, the bound lies at $E = \lfloor 50 \cdot (\ln(100) + \ln(\ln(100)) \rfloor = 307$ edges. Instances with 100 vertices and more than 307 edges are mostly yes-instances, containing one or more Hamiltonian cycles. Instances with fewer than 307 edges are overwhelmingly no-instances, not containing even a single Hamiltonian cycle. For many algorithms ofor the problem, such as Martello's and Rubin's only instances with 100 vertices and close to 307 edges are hard, taking superpolynomial time – all others are easy. Although terminology such as "hard", "easy"," and "close to" lacks rigorous quantification,compelling evidence can be found in a large number of highly cited empirical studies on the topic (e.g. [1],[13],[15]). The main point is this: for the undirected Hamiltonian cycle problem, there is no known polynomial-time algorithm for the worst-case instance. But many instances can nevertheless be solved in polynomial time, and the problem's order parameter, the number of edges, roughly indicates the location of easy instances. Finally, the no/yes phase transition for this problem is firmly rooted in theory, as the probability fthat a randomly generated graph is Hamiltoniand depending on the number of vertices, can be calculated exactly [14]S. till, iit should benoted that this is a limit theorem for $v \to \infty$, and empirical results may differ (substantially) for smaller problem instances.

II. SATISFIABILITY

Satisfiability, another NP-complete problem, asks whether a given formula of Boolean variables and operators has an assignment that evaluates the entire formula to true. The canonical formulation known as CNF3SAT has at most 3 variables per clause and is written in conjunctive normal form, meaning that ∨ and ¬ can appear inside clauses, with ∧ between them. As a running example for this paper

$$(x_1 \vee x_2 \vee \neg x_3) \wedge (x_3 \vee \neg x_2 \vee x_4) \quad (2)$$

is a CNF3SAT-instance with four variables x_1, x_2, x_3 and x_4 nested in two clauses enclosed by brackets we will write $nVar = 4$ and $nClauses = 2$. This instance has multiple satisfying assignments, among which [x_1:= true, x_2:= true, x_3:= true, x_4:= false], and is thereby a yes-instance, even though it also has falsifying assignments such as [x_1:= true, x_2:= true, x_3:= false, x_4:= false]; the fact that it has at least one satisfying assignment makes it a yes-instance.

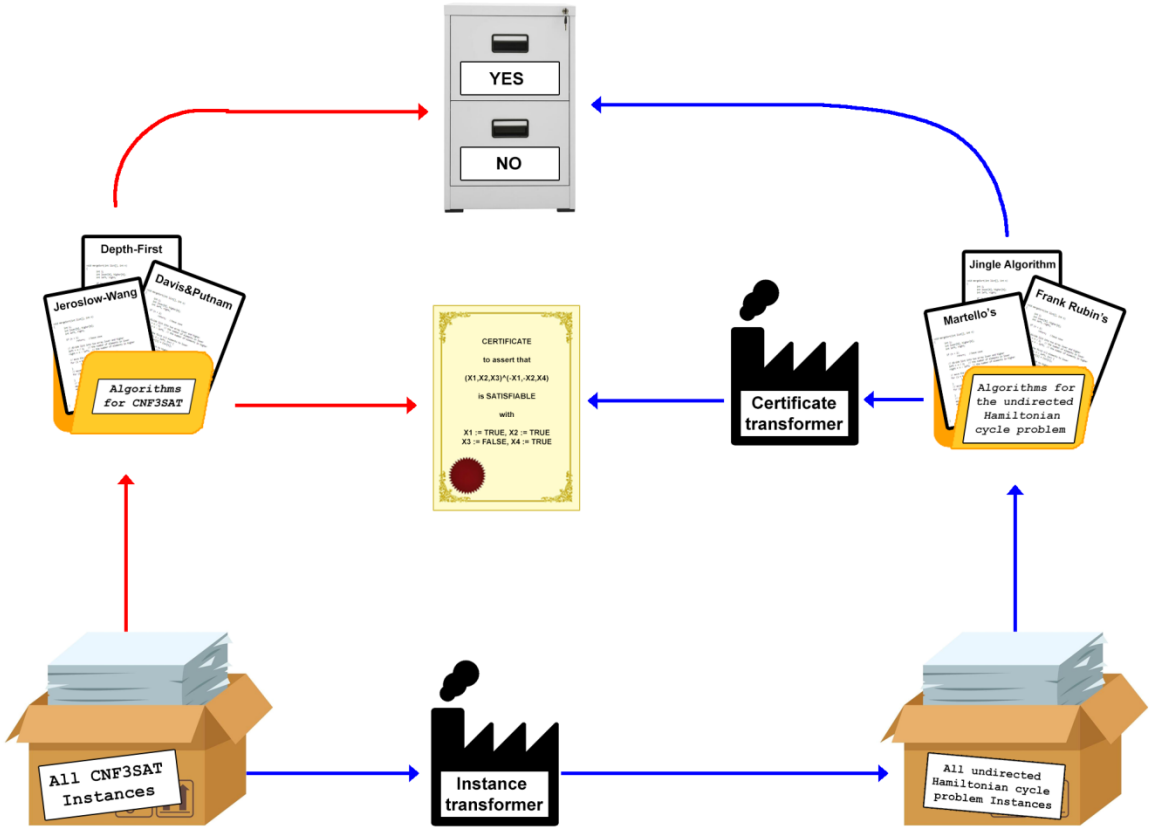


Figure 1. Two routes of solving a CNF3SAT problem instance. Red arrows: picking an exact algorithm directly suitable for the problem always gives a 100% correct yes/no answer to the problem instance, as well as a certificate of the solution (the actual variable assignments) if it exists. Blue arrows: alternatively, the CNF3SAT problem instance can be transformed into an instance of the undirected Hamiltonian cycle problem, for which a different portfolio of exact algorithms with different strengths and weaknesses is available. Even though the yes/no outcome of such an algorithm on the transformed instance is the same, the runtime could be different. Not all transformers are immediately supplied with a certificate transformer, but the one in this paper is.

At the time of writing, the fastest exact algorithm reported in the literature appears to be the PPSZ, with a complexity of O(1.307n) for n variables, which is exponential, as we expect for an NP-complete problem. Just like the undirected Hamiltonian cycle problem, CNF3SAT also exhibits a phase transition, and, again, exact algorithms exist whose runtime depends on the distance to the phase transition. For this problem, the order parameter, indicative of its instance hardness, is the ratio of clauses to variables (σ), and its anchor point lies close to

$$\sigma \approx 4.26 \qquad (3)$$

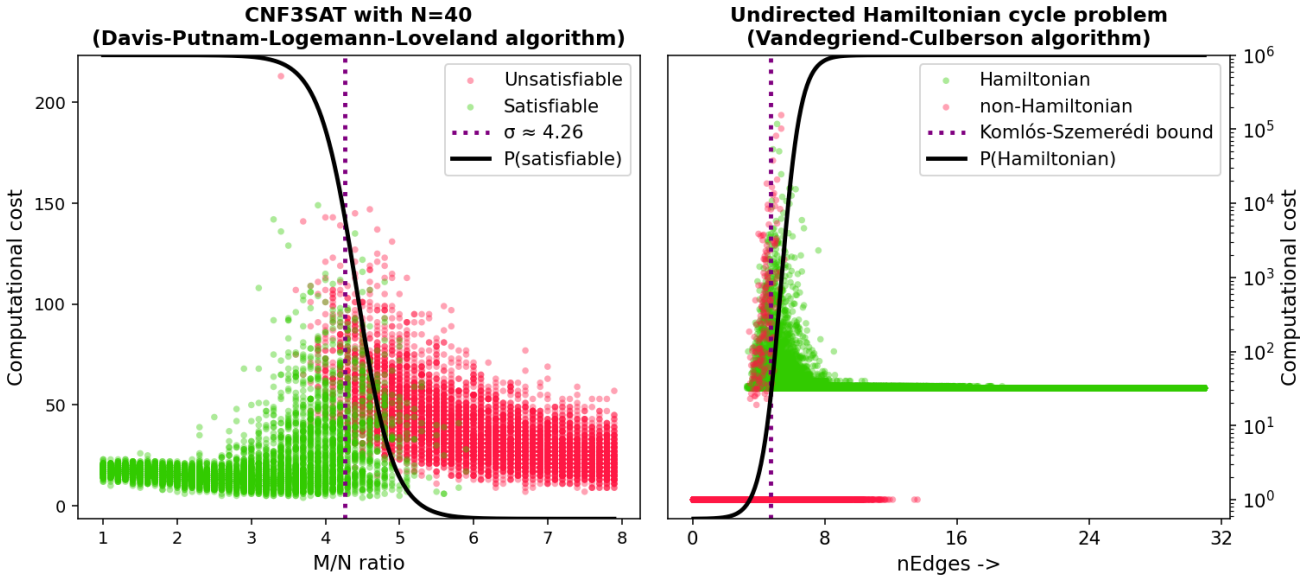


Figure 2. Many NP-complete problems have an 'order parameter' (on the horizontal axis) that indicates where the really hard problem instances lie. More often than not, these hardest instances are found near the anchor point of the 'phase transition' that separates the yes-instances from the no-instances (left) or vice versa (right).

This means that an instance with 10 variables and fewer than 42 clauses is likely a (satisfiable) yes-instance. If it has more than 43 clauses, it is likely a no-instance, for which not a single satisfying assignment exists[1].

The acute reader might have noticed that this phase transition, although phenomenologically similar, is also *inverse* to the undirected Hamiltonian cycle problem's phase transition. CNF3SAT's phase transition is a yes/no phase transition, whereas the undirected Hamiltonian cycle problem has a no/yes phase transition. The reason for this reversal is the natural way of counting and the direction of the English language; the minimum possible value for the order parameter (zero edges for UHCP or just one clause for CNF3SAT, respectively) should conventionally be on the left end of a horizontal axis. This minimum value falls in the no-region of the undirected Hamiltonian cycle problem but in the yes-region of CNF3SAT (Figure 2). The right-hand side of the order parameter axis is the open side, where the maximum value of the order parameter (number of edges for Hamiltonian cycle, number of clauses for CNF3SAT) increases with instance size. For the undirected Hamiltonian cycle problem, the minimum value of its order parameter is zero edges, a graph consisting of only loose vertices, and of course, non-Hamiltonian. The maximum value of its order parameter is $maxEdges = ½v \cdot (v - 1)$, corresponding to a fully connected graph and notably, *maxEdges* grows quadratically with the instance size v. It is interesting to realize that the anchor point of its no/yes phase transition however, grows much slower, namely as $½v \cdot (ln(v) + ln(ln(v)))$ and therefore, as v increases, the percentage of yes-instances approaches 100%. In the limit, all graphs of size v are Hamiltonian.

[1] values for *nClauses* are rounded off for convenience

For CNF3SAT, the opposite is true. The minimum value of its order parameter is 1, a formula with 1 clause and 3 variables that is always satisfiable. Assuming that every CNF3SAT clause is unique and has 3 distinct variables, with $2^3 = 8$ options to place negations, the order parameter's maximum value is

$$maxClauses = \frac{8nVar!}{3!(nVar-3)!} \quad (4)$$

a number that grows polynomially as $\approx (nVar)^3$. The anchor point of the yes/no phase transition, however, grows only linearly as $\approx 4.26n$. This means that as $nVar$ grows, the fraction of no-instances approaches 100%, the exact opposite of the undirected Hamiltonian cycle problem. Yet we are about to show that *every* instance of CNF3SAT can be transformed into an equivalent undirected Hamiltonian cycle problem, a paradox so bizarre it needs a name, especially considering that the inverse reduction (reducing UHCP to CNF3SAT) is also possible. But it gets even weirder.

III. REDUCING CNF3SAT to UNHCP

NP-complete problems are the hardest problems in NP. The common way of proving a problem P_2 is NP-complete is by taking some problem P_1, which is already known to be NP-complete, and providing a function that can transform every instance of P_1 to an instance of P_2 (Figure 1). Once such a transformation has been provided, one can safely conclude that P_2 is at least as hard as P_1, classifying P_2 as NP-complete as well. In such a case, we say "P_1 reduces to P_2," also written as $P_1 \leq P_2$ in this paper[2]. Reductions of this type are transitive, meaning that if $P_1 \leq P_2$ and $P_2 \leq P_3$, then $P_1 \leq P_3$ as well. As such, the entire class of NP-complete problems can be seen as an immense tree of problems interconnected by reduction sequences with SAT (or CircuitSAT) at its root (visualized on page 96 in [9]).

In this paper, we will elaborate on the sequence $CNF3SAT \leq DHCP \leq UHCP$, but this is not the only possible reduction sequence showing the problem's NP-completeness. In fact, Richard Karp's seminal paper itself exhibits a different sequence, namely $CircuitSAT \leq SAT \leq Clique \leq VertexCover \leq DHCP \leq UHCP$ [9]. It is far from trivial to determine whether this reduction sequence will have the same, similar, dissimilar, or completely different effects on the instance size and order parameter values of the resulting UHCP instances. A third reduction, using two types of 'devices' (undirected subgraph patterns), is given by Giri Narasimhan without citing a source, suggesting he might have constructed the reduction himself [16]. A particularly interesting aspect of this reduction is that it is a direct reduction from CNF3SAT to the UHCP, which warrants further investigation. The lack of a formal, thoroughly peer-reviewed source that gives a *complete* explanation of the reductions is problematic and raises questions. It appears that correctness is either communally accepted or remains unquestioned. For this reason, we still mention these informal (but usually largely correct) examples, but we also explicitly perform the reduction ourselves in Sections 3, 4, 5, and 6.

There might be other possible reduction sequences from CircuitSAT to UHCP, each possibly with its own inflationary or even deflationary properties. Still, in this paper, we will discuss the 2-step reduction CNF3SAT $\leq$ DHCP $\leq$ UHCP from CNF3SAT to the undirected Hamiltonian cycle problem. The second step, from the directed Hamiltonian cycle problem (DHCP) to the undirected Hamiltonian cycle problem, is given by Richard Karp himself in two very cryptic lines of text [9], but accredited without reference to [Robert] Tarjan, author of many related studies [17],[8],[18]. The origins of the first reduction in the sequence are somewhat mysterious. Although it appears in Michael Sipser's textbook in Chapter 7.5 [19], there is no reference to a seminal paper. Still, the proof is widely used, e.g., in a course by Adrian She [20], [21],[22], but most notably in an excellent

[2] Notations differ throughout literature



video by "Professor Bryce" [23], who is almost surely Bryce Wiedenbeck from Davidson College in North Carolina, USA, and who appears usually occupied with game theory [24],[25]. His video was a strong source of inspiration for our study, along with a paper by Filar, Haythorpe, and Taylor [26].

It will be these two reductions, *CNF*3*SAT* ≤ *DHCP* and *DHCP* ≤ *UHCP, that we will inspect in the upcoming sections and deduce their exact inflationary effects* on instance size and order parameters. In this context, the word "reduction" is particularly unfortunate, especially for readers new to the subject. Semantically speaking, "P_1 reduces to P_2" pertains to the overall hardness of both problems in terms of computational complexity classes, i.e., "P_1 reduces to P_2" means "proving that the complexity of problem P_1 is lower than, or at most equal to, the complexity of problem P_2". But when it comes to instance sizes, there is nothing to be reduced. Quite opposite in fact, as such NP-complete reductions (or: transformations) tend to inflate the resulting problem's instance size. Furthermore, it can scatter the resulting problem instances' order parameters over a wide range of values, which could significantly impact computational hardness. For this reason, the phenomenon was colloquially dubbed the "TommyGun phenomenon" during a joint lecture on differential information theory [27] at the ILLC in Amsterdam.

Theorem 1. ***("TommyGun")****: When instances of an NP-complete problem* P_1 *of sizes* n_1 *with order parameter values* p_1 *are by reduction transformed to instances of another NP-complete problem* P_2*, it is possible that not all resulting* P_2*-instances are of size* $n_1 + c$*, where* $c \in Z$*. Furthermore, the relative distance of the* p_1*s to* P_1*'s phase transition might also not be equal to the distance of the* p_2 *to* P_2*'s phase transition.*

In other words, NP-complete reductions can flare out instances over a wide range of sizes and order parameters. This could influence hardness in several ways, but for now, we will examine the flare-out effects in reducing CNF3SAT to the undirected Hamiltonian cycle problem.

IV. REQUIREMENTS FOR THE REDUCTIONS

In the first reduction, we will show that for each CNF3SAT formula, we can construct an equivalent instance of the *directed* Hamiltonian cycle problem (DHCP) such that if and only if the CNF3SAT instance is a yes-instance, the corresponding DHCP instance is also a yes-instance. A few preemptive sidenotes are important here:
1. The reduction (or: transformation) from the CNF3SAT-instance to the DHCP-instance is systematic and general. There are potentially infinitely many CNF3SAT instances, simply because the number of variables in the formula can be as large as desired; there is no upper bound. The systematic transformation that a computer program can execute converts any given CNF3SAT instance to a corresponding equivalent DHCP instance.
2. The transformation can be done in polynomial time. By the end of this Section, this will be obvious to the reader.
3. We do not need to know a priori if the CNF3SAT instance is a yes-instance (has a true assignment). We can first convert it to a DHCP instance and then resolve it. If the instance turns out to be non-Hamiltonian, then the corresponding CNF3SAT instance is not satisfiable, and vice versa.
4. For true NP-completeness, there must also be a polynomial-time verification program that verifies whether a given certificate (or "solution") is actually a yes-result. For both CNF3SAT and the Hamiltonian cycle problem (directed and undirected), such a program exists and is easily constructed with a single FOR loop. We will omit these polynomial-time verification programs from the reductions given in the upcoming sections.

In the second reduction, we show that every DHCP instance can be transformed into a UHCP instance, for which all of the above points also apply. There is an extra thought: if a CNF3SAT instance is transformed into a DHCP instance, and that turns out to be Hamiltonian, a solution (Hamiltonian cycle) can be transformed *back*

into a satisfying assignment for CNF3SAT. This backtransformation of a solution (or, more generally, a "certificate") is not typical of NP-complete reductions (see, e.g., textbooks [28] or [19]) but is nonetheless a valuable asset that can make a theoretical-existential reduction into a useful practical instrument.

V. THE FIRST REDUCTION

The first reduction transforms a CNF3SAT Instance into an instance of the directed Hamiltonian cycle problem (DHCP). We will use our running example at each step for clarity. Still, please remember that the reduction is general: it can transform any given CNF3SAT Instance into a DHCP instance such that the instances are yes/no equivalent. Our running example CNF3SAT-instance $(x_1 \vee x_2 \vee \neg x_3) \wedge (x_3 \vee \neg x_2 \vee x_4)$ has $nVar = 4$ variables in $nClauses = 2$ clauses.

Step 1: For variable $x1$, arrange $(2 \cdot nClauses) + 2$ vertices on a horizontal line, and connect every vertex to its right-hand neighbor with a directed edge, and link every vertex to its left-hand neighbor with a directed edge (Figure 3). The vertices now form a bi-directional chain with 2 corner vertices and $2 \cdot nClauses$ internal vertices. Label the leftmost corner vertex "TRUE →" and the rightmost corner vertex "← FALSE". Label the row as a whole "x_1". These labels and spatial arrangement don't substantially change the DHCP instance we are transforming; they are meant to facilitate understanding and enable the backtransformation of a DHCP certificate to a CNF3SAT certificate later.

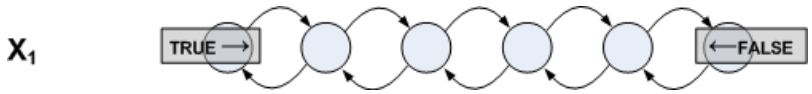


Figure 3. Step 1: Arrange (2 · nClauses) + 2 vertices on a line and connect them to their left and right-hand neighbors, forming a chain.

Step 2: Do the same for all other variables in the CNF3SAT instance, in our case, x_2, x_3, and x_4. We now have $nVar$ chains of $(2 \cdot nClauses) + 2$ vertices for each variable (Figure 4). For our example, that means 4 chains of 6 vertices. For some other formula with $nVar = 7$ variables and $nClauses = 6$, we would end up with 7 chains of 14 vertices.

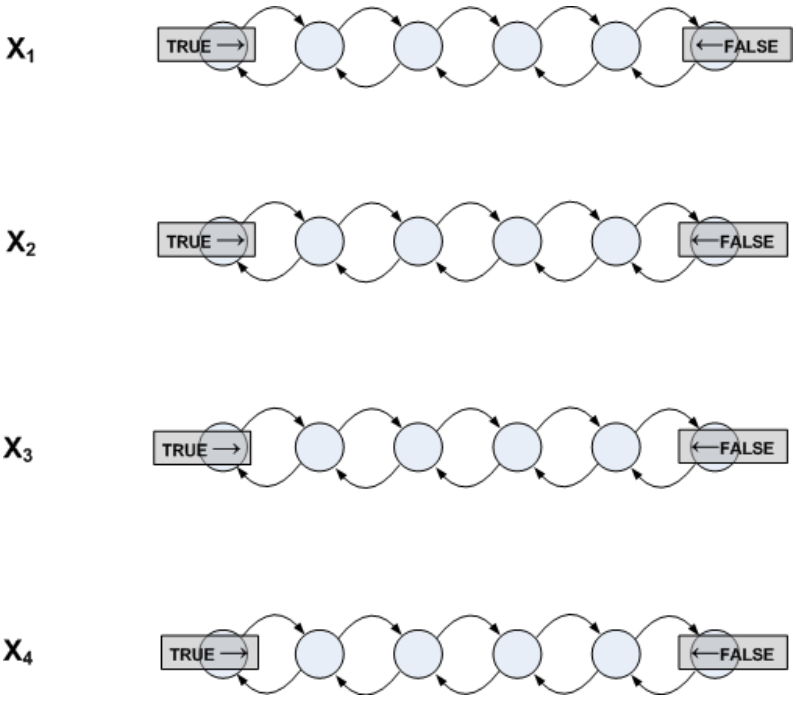


Figure 4. Step 2: Make chains for all variables in the CNF3SAT instance.

Step 3: Connect every corner vertex of each chain to both corner vertices of the next chain, with edges we'll call "inter-chain edges" (Figure 5). For the last chain (x_4 in our case), the interconnections loop back to the first chain (x_1 in our case). From this point onward, the graph is connected (it has no isolated vertices) and is also Hamiltonian. There are, in fact, $2^4 = 16$ directed Hamiltonian cycles in this graph. Each chain can be traversed either left-to-right or right-to-left, in any arbitrary combination following the inter-chain edges. Thus, we could transform every Hamiltonian cycle (or certificate) from this instance back into a variable assignment for the CNF3SAT instances, following the labels we applied in steps 1 and 2. The important discrepancy, however, is that not all Hamiltonian cycles in this graph correspond to a satisfying assignment ... yet. So in its current state, the transformation is not yet a valid reduction, but we'll solve this in the next step.

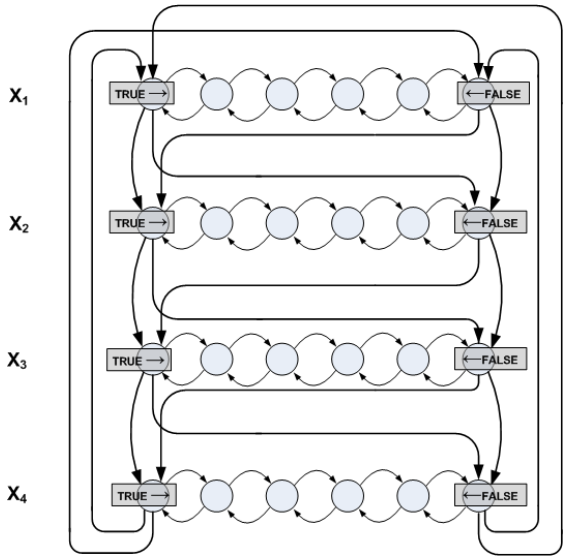


Figure 5. Step 3: Make 'interchain connections' from both corner vertices of a chain to both corner vertices of the next chain.

Step 4: Add one vertex to the graph for each clause in the CNF3SAT formula. In our case, 2 vertices in for 2 clauses. Now, in each chain, the first two internal vertices correspond to Clause 1, the second two internal vertices correspond to Clause 2. An edge is drawn from one of the corresponding internal vertices to the clause, and back to the other of the two, and the direction corresponds to the chain's variable's value that satisfies a given clause. As an example, in our formula $(x1 \vee x2 \vee \neg x3) \wedge (x3 \vee \neg x2 \vee x4)$, variable $x3$ can satisfy *Clause*2 when assigned true, so we connect from the 3rd internal vertex to *Clause*2 and back from *Clause*2 to the 4th internal vertex. Notice how this creates an 'excursion' to *Clause 2 in the true direction* of chain $x3$ (Figure 6). But variable $x3$ can also satisfy *Clause 1 if it is assigned the value false, so we also create an excursion to Clause 2 from chain x3, but this time in the false-direction, emanating from the 2nd internal vertex to Clause 2* and then back to the 1st internal vertex of chain $x3$. Notice again, how the 'satisfactory direction' corresponds to the edge direction: *Clause 2* can only be visited from $x3$ in the false-direction, just as the second clause in our running example can only be satisfied by $x3$ when it is set to false.

Our reduction is now complete. We can now transform any CNF3SAT Instance into a DHCP instance and use any DHCP algorithm to solve it. If the DHCP instance is Hamiltonian, then our CNF3SAT instance is satisfiable, but if the DHCP instance does not contain a Hamiltonian cycle, a satisfying assignment for the CNF3SAT instance does not exist. And as a bonus, a returned Hamiltonian cycle can be transformed back to a satisfying CNF3SAT-assignment by using the labels.

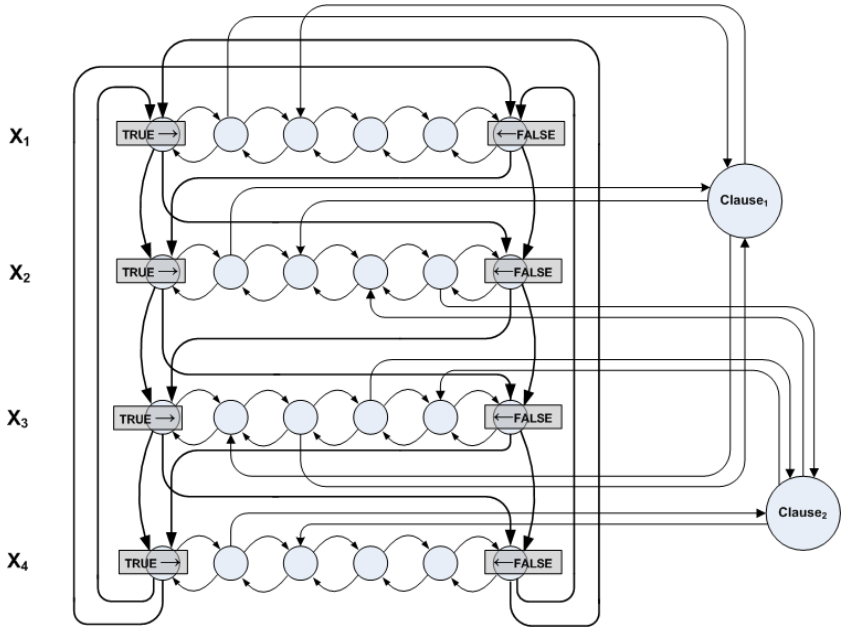


Figure 6. Step 4: A complete reduction from the CNF3SAT instance $(x_1 \vee x_2 \vee \neg x_3) \wedge (x_3 \vee \neg x_2 \vee x_4)$ to an instance of the directed Hamiltonian cycle problem (DHCP).

VI. INFLATIONARY PROPERTIES AND MALLEABILITY OF NUMBERS

Here we arrive at the main topic of this paper: are the resulting DHCP instances harder or easier than the original CNF3SAT instances? We will make a preliminary assessment by quantifying the inflationary effect of the reduction. For this transformation from aCNF3SAT-instance

1. We need $(2 \cdot nClauses + 2)$ vertices for one chain.
2. So for all $nVar$ chains together, we need $nVar \cdot (2 \cdot nClauses + 2)$ vertices.
3. We need one vertex for each clause, so $nClauses$ vertices for all clause vertices.
4. We need $2 \cdot (2 \cdot nClauses + 1)$ edges for each chain.
5. So for all chains, we need $nVar \cdot (2 \cdot (2 \cdot nClauses + 1))$ edges.
6. We need $4 \cdot nVar$ inter-chain edges, namely 4 for each chain to the next chain.
7. We need 6 edges per clause, connecting them to the chains, so $6 \cdot nClauses$

This means that for a CNF3SAT instance of $nVar$ variables and $nClauses$ clauses, the resulting DHCP instance will have

$$\text{nVertices} = nVar \cdot (2 \cdot nClauses + 2) + nClauses$$

$$\text{nEdges} = nVar \cdot (2 \cdot (2 \cdot nClauses + 1)) + (4 \cdot nVar) + 6 \cdot nClauses$$

meaning that for our running example $(x_1 \vee x_2 \vee \neg x_3) \wedge (x_3 \vee \neg x_2 \vee x_4)$, having $nVar = 4$ and $nClauses = 2$, the resulting DHCP instance has $nVertices = 4 \cdot (2 \cdot 2 + 2) + 2 = 26$ vertices, and $nEdges = 68$ edges. Luckily, the blowup from CNF3SAT to DHCP is itself linear, but this does not imply ease. For example, a relatively small CNF3SAT instance with 21 variables and 377 clauses, quite doable for most algorithms, yields a UHCP



instance with 16,253 vertices and 34,056 edges – far beyond any measure of modern-day computability. These numbers, however, are somewhat malleable.

The numbers and equations given in the enumerated list in Section 5.1 pertain to the most default form of the reduction, and leave some room for variation. First of all, point 1 is an "at most" number. Notice that in the x_1-chain of our example, the 3rd and 4th vertices could be removed by connecting the 3rd vertex immediately to the rightmost corner vertex without changing the Hamiltonicity of the instance. In fact, in all chains, the pairs of vertices not connecting to their corresponding clauses can be removed. For very large *nVar* and *nClauses*, one can expect a lot of such redundancy, because not every variable can be in every clause of a CNF3SAT instance, which has at most 3 variables per clause. Note that this redundancy, if applied, would also decrease the number of edges in a chain, making point 3 an "at most" value as well. Second, *nVar* should be considered an *actual* value, not an *intended* value. That means that if the transformable CNF3SAT instance is randomly generated by assigning *nVar* randomly chosen variables to *nClauses* (randomly chosen) clauses, all *nVar* variables should be chosen at least once. Conversely, one could state that *nVar* is the number of variables in the CNF3SAT Instance after its generation, regardless of the originally intended number of variables. Of course, this all depends on the algorithmics of the random CNF3SAT instance generator. In any case, *nVar* is the *actual* number of variables in the formula. Third, the number of clauses in the CNF3SAT instance is manipulable. One could introduce duplicate clauses that are redundant and irrelevant to the satisfiability of the instance, thereby artificially inflating the resulting DHCP instance and likely increasing the runtime of the DHCP algorithm. Note that for this paper, we assume every clause to be unique (also see Equation 4). Fourth, the DHCP instance can be almost arbitrarily inflated without affecting its Hamiltonicity. One could add vertices in a chain for clauses that do not exist. It will slightly influence the instance's hardness for UHCP algorithms (at approximately one recursion per added vertex). Still, its size increases greatly, so one could say its relative hardness actually decreases when inflated. Also, some versions of the reduction, like Adrian She's, use start and end vertices [20]. An arbitrary number of vertices could be placed between these without fundamentally changing their Hamiltonicity. Bryce Wiedenbeck's reduction [23] omits these vertices, as does ours.

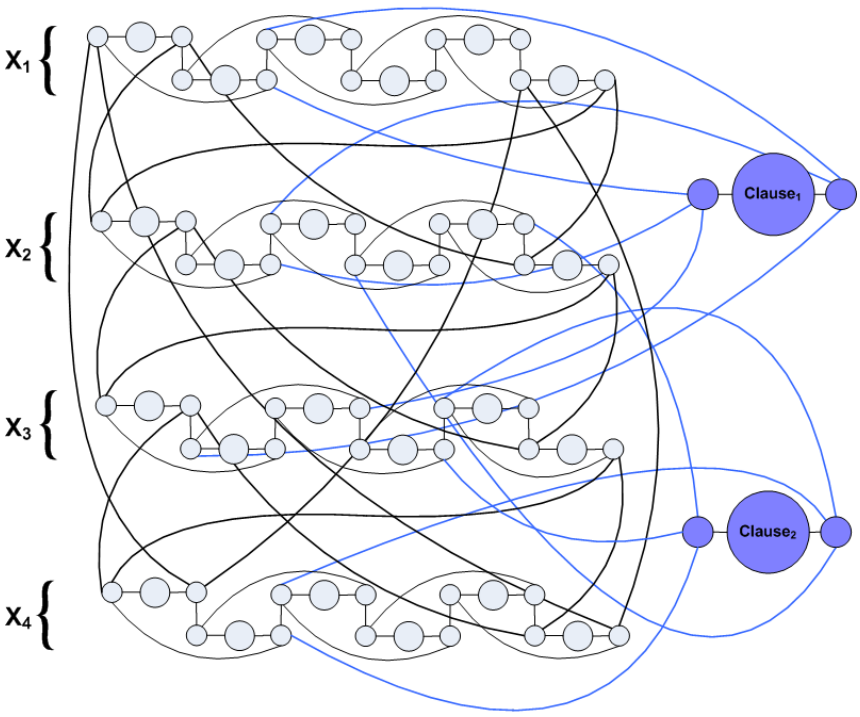


Figure 7. Any instance of the directed Hamiltonian cycle problem can be transformed into an undirected Hamiltonian cycle problem instance by replacing each vertex Vx of the DHCP instance with a triplet Vx: in, Vx: mid and Vx: out in an undirected Hamiltonian cycle problem instance, whereafter a directed edge (Vx, Vy) can be transformed to an undirected edge (Vx: out, Vx: in). As such, these points show that the resulting instance sizes and order parameter values are not rock-hard but malleable if one wants them to be. The instance size increase from CNF3SAT to DHCP is linear in this default case, but it could artificially be pumped up to any size, e.g., "to make an argument", by adding extra vertices and edges. Conversely, its size might decrease slightly by removing redundancy in the instance, the extent of which depends on the particular CNF3SAT instance that was transformed.

VII. THE SECOND REDUCTION

The NP-completeness of the *undirected* Hamiltonian cycle problem comes from a succinct description by Richard Karp himself [9] and reduces from the *directed* Hamiltonian cycle problem. Luckily, the procedure is relatively straightforward, and works as follows: for every vertex *Vi* in the directed graph G_{dir}, create three vertices *Vi*: *in*, *Vi*: *mid,* and *Vi*: *out* in the undirected graph G_{undir}. Connect *Vi*: *in* and *Vi*: *out* to *Vi*: *mid* with undirected edges. Then, for every directed edge (*Vi, Vj*) of G_{dir}, create an undirected edge from *Vi*: *out* to *Vj*: *in* (see Figure 7).

The resulting graph*, Gundir,* will have the same Hamiltonicity as G_{dir} and the same order of vertex visits. When the labeling (omitted in Figure 7 due to space constraints) is fixed as suggested (Vi to Vi: in, Vi: mid, and *Vi*: *out*), a Hamiltonian cycle in G_{undir} can be transformed back to G_{dir}, hence, a certifcate-backtransformation also exists. One slight complication, however, is that the UCHP solver could return a Hamiltonian cycle that is the direct inverse of a corresponding Hamiltonian cycle in G_{dir}. Assuming we don't touch the UHCP solving algorithms, which seems prudent, the certificate backtransformer should check whether the certificate should be

read forward or backward, which can be done in polynomial time. This second transformation again inflates the instance. For a DHCP-instance to a UHCP-instance, the numerics are easy: a directed graph $G_{dir}(v, e)$ with v vertices and e edges projects to an undirected graph of $G_{undir}(3v, e+2ve)$ vertices and edges. A relatively modest blowup, even though a factor of 3 is not insignificant, considering that all algorithms for NP-complete problems are of superpolynomial complexity. For a transformation of CNF3SAT directly to UHCP, the numerics are a bit more elaborate.

1. We now need $(6 \cdot nClauses + 6)$ vertices for one chain.
2. So for all $nVar$ chains together, we need $nVar \cdot (6 \cdot nClauses + 6)$ vertices.
3. We now need 3 vertices for each clause, so $3 \cdot nClauses$ vertices for all clause vertices.
4. We now need $2 \cdot (2 \cdot nClauses + 2)$ inter-triplet edges in each chain.
5. We still need $2 \cdot (2 \cdot nClauses + 1)$ edges for each chain, connecting the triplets.
6. So for all chains, we need $nVar \cdot (2 \cdot (2 \cdot nClauses + 2) + 2 \cdot (2 \cdot nClauses + 1))$ edges.
7. We still need $4 \cdot nVar$ inter-chain edges, no changes here.
8. We need $8 \cdot nClauses$ edges for each clause triplet, of which 2 are inter-triplet.

Summarizing, this means that for a CNF3SAT instance of $nVar$ variables and $nClauses$ clauses, the resulting UHCP instance will have

$$nVertices = nVar \cdot (6 \cdot nClauses + 6) + 3 \cdot nClauses.$$

$$nEdges = nVar \cdot (2 \cdot (2 \cdot nClauses + 2) + 2 \cdot (2 \cdot nClauses + 1)) + 4 \cdot nVar + 8 \cdot nClauses$$

VIII. IMPLICATION AND DISCUSSION

The numerical analysis in the previous section yields some puzzling implications. As a numerical exploration, we perform reductions for CNF3SAT formulas with $nVar = \{4, 10, 20, 50, 100, 200, 500, 1000\}$. For all these $nVar$ numbers, the minimal clause numbers, trivially determined as $\lfloor \frac{nVar}{3} \rfloor$, maximal clause numbers (as given by Equation 4), and intermediate clause numbers (halfway between the minimum and the maximum) are given in Table 1. As an immediate result, the instance size after reduction (*nVertices*) grows substantially after the NP-complete reduction. Counterintuitively, this is due more to the explosion of *maxClauses* than to the inflationary effects of the reduction itself. This results in a single value of *nVar*, the key parameter in CNF3SAT's complexity, which scales to a large number of *nVertices*. Even though the inflation is non-constant and "only polynomial," these numbers are quite significant from a computational complexity standpoint. As an example, for a relatively modest 1000-variable CNF3SAT instance, the resulting UHCP instance size ranges between 2.01 · 106 and 7.98 · 1012 vertices, which are almost a factor of 4 million apart (Table 1), an insurmountable spread for any exact algorithm designed to solve problems, NP-complete or otherwise. But as a secondary result, the location of the resulting UHCP problem instances relative to the no/yes phase transition is also nontrivial. Comparing the 4th and 5th columns of Table 1 shows how the number of edges for an instance, and the number of edges at the Komlós-Szemerédi bound. Due to the edge sparsity of the resulting UHCP instances, they are located far *before* the threshold function's anchor point. This is weird. The area before the Komlós-Szemerédi bound is the area where nearly all UHCP instances are non-Hamiltonian. Yet *all* CNF3SAT instances, both satisfiable and unsatisfiable, project to this no-region of UHCP. How to understand this apparent contradiction? An answer might partly come from the UHCP's pinnacle of hardness, which is itself paradoxical and contradictory. The fact that the hardest instances of NP-complete problems lie around their yes/no phase transition has been known for decades. Still, a recent study showed that the *uniquely* hardest instance lies far beyond the Komlós-Szemerédi bound, a no-instance in an area so dense with yes-instances it could almost not

exist [29],[30],[31]. It is a single superhard no-instance amidst an ocean of trivially easy yes-instances. But paradoxically enough, even though this superhard instance *itself* is the hardest instance in existence, the *average* hardest instances are still found along the Komlós-Szemerédi bound. So how is it possible that despite the vast body of literature, this instance had never been found? The answer is given by *structure*. The hardest instance of UHCP, situated in an overwhelmingly easy region of the problem, is a highly structured graph. And despite most historical hardness studies involving large numbers of randomly generated instances, such structured instances were never found. This can be explained by an argument based on Kolmogorov complexity, which affirms that the chance of randomly generating a structured instance is astronomically small [32]. A similar argument might be surfacing from the paradoxical Tommy Gun effects of the reductions made above. How can all the CNF3SAT instances, both satisfiable and unsatisfiable, project solely onto the no-area of the UHCP? The answer could again be structure. It is the recipe, the general procedure of these NP-complete reductions, that produces structured instances. By the same argument from Kolmogorov complexity, these instances will almost certainly not turn up in randomized instance sets – perhaps not even in some structured instance sets. So the problem with structured instances appears to be: you don't know they exist until you know they exist, and they don't play by the rules. Possibly, the hardness of structured instances in general is completely different from that of unstructured instances. Table 1: The "TommyGunning" effect of reducing CNF3SAT of size nVar to UHCP gives a range of instance sizes (nVertices, 3rd column) as well as order parameter values (nEdges, 4th column). For this particular reduction, the distance to the phase transition (KS-bound, 5th column) also changes substantially (nEdges bölü payda, KS, eksi bound son payda, 6th column).

nVar	**nClauses**	**nVertices**	**nEdges**	**KS-bound**	**nEdges /KS-bound**
4	2	78	112	227	0.4934
4	17	483	772	1932	0.3996
4	32	888	1432	3865	0.3705
10	4	312	428	1169	0.3661
10	482	30426	44404	192558	0.2306
10	960	60540	88380	405919	0.2177
20	7	981	1324	4325	0.3061
20	4564	561492	785128	4441815	0.1768
20	9120	1121880	1568760	9291746	0.1688
50	17	5451	7304	29315	0.2492
50	78409	23758227	32304808	235392690	0.1372
50	156800	47510700	64601900	488141552	0.1323
100	34	21102	28208	129307	0.2181
100	646817	390031251	525216004	4439822572	0.1183
100	1293600	780041400	1050403800	9163167865	0.1146
200	67	81801	109204	561888	0.1944
200	5253634	6320122902	8468859208	81161417271	0.1043
200	10507200	12640162800	16937607600	1.66894E+11	0.1015
500	167	504501	673004	3961929	0.1699
500	82834084	2.48751E+11	3.3233E+11	3.66994E+12	0.0906
500	165668000	4.97501E+11	6.6466E+11	7.51879E+12	0.0884
1000	334	2011002	2682008	17283825	0.1552
1000	664668167	3.99E+12	5.32532E+12	6.46034E+13	0.0824
1000	1329336000	7.98E+12	1.06506E+13	1.32067E+14	0.0806

This observation opens up many new perspectives. Is it possible for some reduction to project both yes- and no-instances of one NP-complete problem solely into the yes-region of another NP-complete problem? Or into both yes and no areas? Maybe. In any case, it is starting to look like random instances form a cloud obfuscating the true nature of NP-completeness, which the contours of their structured instances might define. Even though the phase transitions and associated instance hardnesses are real, the rules might not apply to instances that have structure in the Kolmogorov sense of the word. Sadly, these conclusions also upend all our ideas of solving the P =? NP problem along these routes. We are looking for evidence that hard CNF3SAT instances project into easy regions of the UHCP, which could make it possible to efficiently solve the very hard instances via NP-complete reductions. While it is certainly true that hard CNF3SAT Instances project into an easy region of the undirected Hamiltonian cycle problem, hopes of easy solvability have largely been dashed. The way forward is to solve a large number of transformed CNF3SAT instances using a portfolio of UHCP algorithms. But as these algorithms are general, they may perform well only on random (i.e., Kolmogorov-unstructured) instances, and far less well on structured graphs. The only way to find out is to try and see, as always, in future work.

CREDIT STATEMENT & ACKNOWLEDGEMENTS

Daan van den Berg: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration. **Joeri Sleegers**: Methodology, Software, Validation, Writing - Review & Editing. **Thomas Meeus**: Methodology, Software, Validation, Writing - Review & Editing. **Pieter Adriaans**: Conceptualization, Writing - Review & Editing. The first author thanks Rebecca Reiffenhauser (UvA) and Maarten Stol (VU University, BrainCreators) for their help in discussing various aspects of the subject matter. It was fun, let's stay in touch, everyone.

REFERENCES

[1] Peter Cheeseman, Bob Kanefsky, and William M. Taylor. Where the really hard problems are.In Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1 IJCAI'91, pages 331–337, San Francisco, CA, USA, 1991—Morgan Kaufmann Publishers Inc.URL: http://dl.acm.org/citation.cfm?id=1631171.1631221.

[2] Nikita Sazhinov, Ruben Horn, Pieter Adriaans, and Daan van den Berg. The partition problem, and how the distribution of input bits affects the solving process. In IJCCI, pages 143–150, 2023.

[3] Ruben Horn, Sarah L Thomson, Daan Van Den Berg, and Pieter Adriaans. The easiest hard problem: Now even easier. In Proceedings of the Genetic and Evolutionary Computation Conference Companion, pages 97–98, 2024.

[4] Daan van den Berg and Pieter Adriaans. Subset sum and the distribution of information. In IJCCI, pages 134–140, 2021.

[5] Jiayang Chen, Zhize Wu, Sarah L Thomson, and Thomas Weise. Frequency fitness assignment: Optimization without bias for a good solution outperforms randomized local search on the quadratic assignment problem. 2024.

[6] Sébastien Verel, Sarah L Thomson, and Omar Rifki. Where the really hard quadratic assignment problems are: the QAP-SAT instances. In European Conference on Evolutionary Computation in Combinatorial Optimization (Part of EvoStar), pages 129–145. Springer, 2024

[7] Iris Pijning, Levi Koppenhol, Danny Dijkzeul, Nielis Brouwer, Sarah L Thomson, and Daan van den Berg. The performance of frequency fitness assignment on JSPS for different problem instance sizes. In 16th International Joint Conference on Computational Intelligence, IJCCI 2024, pages j0–260. SciTePress, 2024.

[8] Michael R Garey, David S. Johnson, and R Endre Tarjan. The planar Hamiltonian circuit problem is NP-complete. SIAM Journal on Computing, 5(4):704–714, 1976.

[9] Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, editors, Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Ofce of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department, Boston, MA, 1972. Springer US. doi:10.1007/978-1-4684-2001-2_9.

[10] Frank Rubin. A search procedure for Hamilton paths and circuits. J. ACM, 21(4):576–580, October 1974. URL: http://doi.acm.org/10.1145/321850.321854, doi:10.1145/321850. 321854.

[11] Silvano Martello. Algorithm 595: An enumerative algorithm for finding Hamiltonian circuits in a directed graph. 1983.

[12] Basil Vandegriend and Joseph Culberson. The gn, m phase transition is not hard for the Hamiltonian cycle problem—Journal of Artificial Intelligence Research, 9:219–245, 1998.

[13] Gijs van Horn, Richard Olij, Joeri Sleegers, and Daan van den Berg. A predictive data analytics for the hardness of Hamiltonian cycle problem instances. DATA ANALYTICS 2018: The Seventh International Conference on Data Analytics, 2018.

[14] János Komlós and Endre Szemerédi. Limit distribution for the existence of Hamiltonian cycles in a random graph. Discrete Mathematics, 43(1):55–63, 1983.

[15] Rémi Monasson, Riccardo Zecchina, Scott Kirkpatrick, Bart Selman, and Lidror Troyansky. Determining computational complexity from characteristic 'phase transitions'. Nature, 400(6740):133–137, 1999.

[16] Giri Narasimhan. NP-completeness: The Hamiltonian-cycle problem & the traveling salesperson problem. https://users.cs.fiu.edu/~giri/teach/UoM/7713/f98/yige/yi12. html#HCR. Accessed: September 13th, 2025.

[17] Pierre Rosenstiehl and Robert E Tarjan. Gauss codes, planar hamiltonian graphs, and stack-sortable permutations. Journal of Algorithms, 5(3):375–390, 1984.

[18] George Pólya, Robert E Tarjan, and Donald R Woods. Hamiltonian and Eulerian paths. In Notes on Introductory Combinatorics, pages 157–168. Springer, 2009.

[19] Michael Sipser. Introduction to the theory of computation. ACM Sigact News, 27(1):27–29, 1996.

[20] Adrian She. Hamiltonian path is NP-complete. https://www.cs.toronto.edu/~ashe/ham-path-notes.pdf. Accessed: September 13th, 2025. Contains a few technical typos; the author is informed about this.

[21] Adrian She and Henry Yuen. Unitary property testing lower bounds by polynomials. arXiv preprint arXiv:2210.05885, 2022.

[22] Samantha Dahlberg, Adrian She, and Stephanie van Willigenburg. Schur and e-positivity of trees and cut vertices. arXiv preprint arXiv:1901.02468, 2019.

[23] Bryce Wiedenbeck. The Hamiltonian cycle problem is NP-complete (Algorithm 24). https://www.youtube. com/watch?v=OHnX-R_SBpM. Accessed: September 13th, 2025.

[24] Bryce Wiedenbeck, Ben-Alexander Cassell, and Michael P Wellman. Bootstrap statistics for empirical games. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, pages 597–604, 2014.

[25] Bryce Wiedenbeck, Michael P Wellman, et al. Learning payoffs in large symmetric games. In AAMAS, volume 15, pages 1881–1882, 2015.

[26] Jerzy A Filar, Michael Haythorpe, and Richard Taylor. Linearly-growing reductions of Karp's 21 NP-complete problems. arXiv preprint arXiv:1902.10349, 2019.

[27] Pieter Adriaans. Differential information theory. arXiv preprint arXiv:2c1.04335, 2021

[28] Michael R Garey and David S Johnson. Computers and intractability, volume 29. Wh Freeman, New York, 2002.
[29] Joeri Sleegers and Daan van den Berg. Looking for the hardest Hamiltonian cycle problem instances. In IJCCI, pages 40–48, 2020.
[30] Joeri Sleegers and Daan van den Berg. The hardest Hamiltonian cycle problem instances: the plateau of yes and the cliff of no. SN Computer Science, 3(5):372, 2022.
[31] Joeri Sleegers, Sarah L. Thomson, and Daan van den Berg. Universally hard Hamiltonian cycle problem instances. 2022.
[32] Ming Li, Paul Vitányi, et al. An introduction to Kolmogorov complexity and its applications, Volume 3. Springer, 2008.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Explainable Commit-Level Vulnerability Intelligence in DevSecOps Pipelines: A Systematic Review and Experimental Analysis

Dasun Sandeepa Wilathgamuwa
Faculty of Computing
University of Sri Jayawardanapura
Nugegoda, Sri Lanka
fc110590@foc.sjp.ac.lk
ORCID: 0009-0000-3113-0818

Malshani Heetiarachchi
Department of Interdisciplinary Studies
Faculty of Information Technology
Moratuwa, Sri Lanka
malshanihettiarachchi@sjp.ac.lk
ORCID: 0009-0002-4559-7974

Abstract
DevSecOps has emerged as a critical paradigm for integrating security into continuous integration and continuous deployment (CI/CD) pipelines, driven by the increasing complexity and rapid delivery requirements of modern software systems. Despite growing adoption, existing security mechanisms remain fragmented and misaligned with developer workflows, limiting their practical effectiveness. Recent advances in artificial intelligence (AI), particularly deep learning and large language models, have demonstrated strong potential for automated vulnerability detection. However, these techniques often fail to integrate effectively into real-world DevSecOps pipelines. Most existing approaches operate at the function or file level and provide limited explanation of their predictions, making it difficult for developers to interpret results during code changes and reducing trust in such tools. This paper presents a structured review of AI-based vulnerability detection approaches, with a particular focus on commit-level analysis and explainability within CI/CD environments. The study analyses 40 selected studies published between 2018 and 2025, collected through searches of major digital libraries and indexed sources. The findings show that only a small proportion of the reviewed studies perform commit-level analysis, while most provide little or no explainability. Based on these findings, this work synthesizes a conceptual framework that integrates commit-level change analysis, AI-based vulnerability detection, and explainable security mechanisms within CI/CD pipelines. The framework aims to improve developer trust, enable faster and more informed decision-making, and enhance the practical adoption of AI-driven security in DevSecOps environments.

Keywords: DevSecOps, Commit level vulnerability detection, Explainable AI, CI/CD security, Software security analytics.

I. INTRODUCTION

DevSecOps has emerged as an extension of the DevOps paradigm, embedding security as a continuous and shared responsibility throughout the software development lifecycle. Modern software systems are increasingly delivered through continuous integration and continuous deployment (CI/CD) pipelines, where development, testing, and deployment occur at high speed. In this context, DevSecOps aims to integrate security controls directly into these pipelines, enabling early vulnerability detection without disrupting development velocity. Despite its widespread adoption, security practices within DevSecOps environments remain fragmented across multiple tools and workflows. Existing studies indicate that security mechanisms are often not tightly integrated into a unified development pipeline, thereby limiting their effectiveness and creating challenges for developers in interpreting and acting on security findings [1–5]. To address the scale and complexity of modern CI/CD environments, artificial intelligence (AI) has been widely explored for vulnerability detection. Techniques such as deep learning, graph-based models, and large language models (LLMs) can capture the semantic and structural properties of source code, often outperforming traditional static and dynamic analysis tools in detecting complex

vulnerabilities [6–8]. However, despite these advances, significant limitations remain. Many AI-based approaches operate as black-box systems, offering limited transparency into their decision-making processes. Furthermore, most techniques focus on file-level or function-level analysis, overlooking commit-level changes that are central to CI/CD workflows. This disconnect reduces their practical applicability in real DevSecOps environments. In addition, issues such as high false-positive rates, alert fatigue, and the lack of standardized mechanisms for integrating detection results with workflow context further reduce developers' trust and hinder adoption [9–13]. These limitations highlight a critical gap between current research and the operational needs of DevSecOps. In particular, there is a need for approaches that combine fine-grained commit-level analysis with explainable outputs that can support developer decision-making during code changes. To address this gap, this paper presents a structured review of AI-based vulnerability detection approaches, focusing on explainability and commit-level analysis within DevSecOps pipelines. The study categorizes existing methods, synthesizes key findings, and identifies unresolved challenges. Based on this analysis, the paper further proposes a conceptual framework that integrates commit-level vulnerability detection with explainable security mechanisms, aiming to bridge the gap between research advances and practical DevSecOps requirements [6], [3], [14].

II. REVIEW METHODOLOGY

1. Search Strategy

A literature search was conducted across IEEE Xplore, the ACM Digital Library, SpringerLink, arXiv, and other indexed sources. The search focused on studies published between 2018 and 2025, covering the evolution of deep learning, large language models, and explainable AI in software security. The search used combinations of the following terms: DevSecOps, CI/CD security, vulnerability detection, commit-level analysis, and explainable AI. Boolean operators (AND, OR) were used to combine the terms across the targeted databases.

2. Study Selection Process

The PRISMA-based selection process identified 100 records. After duplicate removal, 90 records remained for title and abstract screening. Fifty studies proceeded to full-text assessment, and 10 were excluded at that stage. A final set of 40 studies was selected for detailed analysis. The selection criteria were applied consistently across the screening stages to maintain relevance to AI-based vulnerability detection, DevSecOps, CI/CD environments, and explainability.

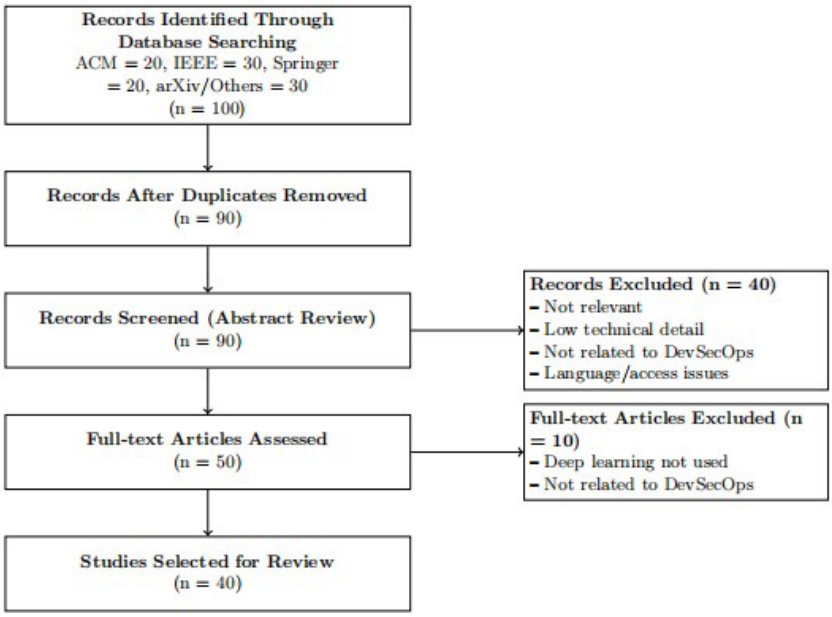


Figure 1. Study selection process following PRISMA methodology

3. Inclusion and Exclusion Criteria

Inclusion criteria were: studies focusing on AI-based vulnerability detection; research related to DevSecOps or CI/CD environments; papers addressing explainability in software security; and peer-reviewed articles or high-quality preprints. Exclusion criteria were: studies unrelated to software security or vulnerability detection; papers lacking technical depth or methodological detail; research not aligned with CI/CD or DevSecOps workflows; and non-English or inaccessible publications

4. Data Extraction and Classification

For each selected study, key attributes were extracted and analyzed, including research category, dataset usage, method type, level of analysis, and explainability support. The extracted data were organized into structured comparison tables to support cross-study analysis and identification of research trends, limitations, and gaps.

III. OVERVIEW OF SELECTED STUDIES

This section presents the key characteristics of the selected literature, including the search sources, keyword categories, and observed patterns in explainability, methodological approaches, and levels of analysis.

Table 1. Distribution of Retrieved Studies by Database

Database	Approx. Papers Retrieved
IEEE Xplore	30
ACM Digital Library	20
SpringerLink	20

arXiv / Others	30
Total	100

Table 2. Search Keywords and Categories

Category	**Keywords**
DevSecOps	DevSecOps, CI/CD security
Detection	Vulnerability detection, software security
AI Methods	Deep learning, machine learning, LLM
Analysis Level	Commit-level, file-level
Explainability	Explainable AI, XAI

1. Search Strategy Summary

Table 1 summarises the distribution of records retrieved from the targeted sources, while Table 2 presents the keyword categories used during the literature search.

2. Analysis of Selected Studies

Figure 2 shows that 79% of the reviewed studies provide no explainability, 11% provide limited explainability, and 10% provide explainable outputs. This indicates that explainability remains largely underexplored despite the increasing adoption of AI-based vulnerability detection techniques.

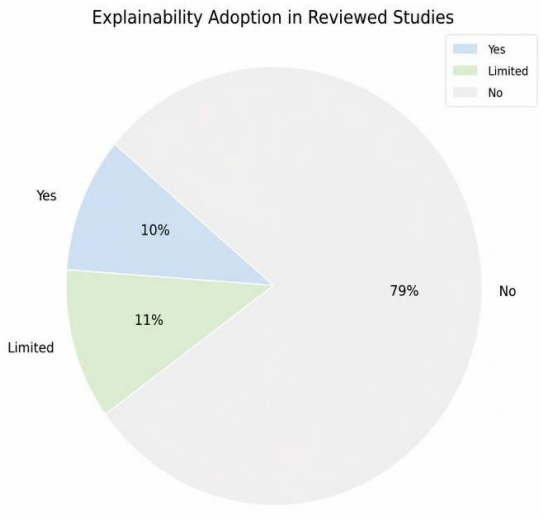


Figure 2. Distribution of explainability adoption across reviewed studies

Figure 3 summarises the methodological approaches represented in the reviewed literature. Survey-based studies are prominent, followed by machine learning and deep learning approaches, while relatively few studies directly explore explainable AI techniques.

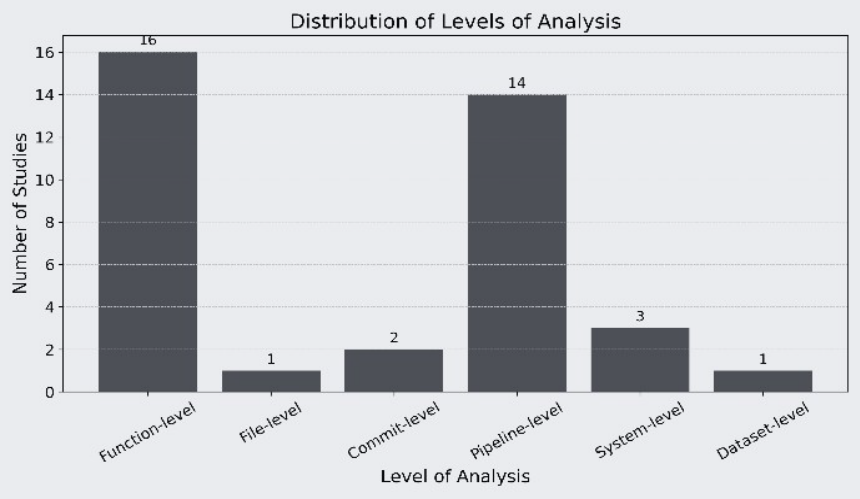


Figure 3. Distribution of methodological approaches in the reviewed literature

Figure 4 presents the distribution of analysis levels. Function-level and pipeline-level analysis dominate, whereas commit-level analysis remains limited. This pattern motivates closer integration of commit-level vulnerability analysis and explainability.

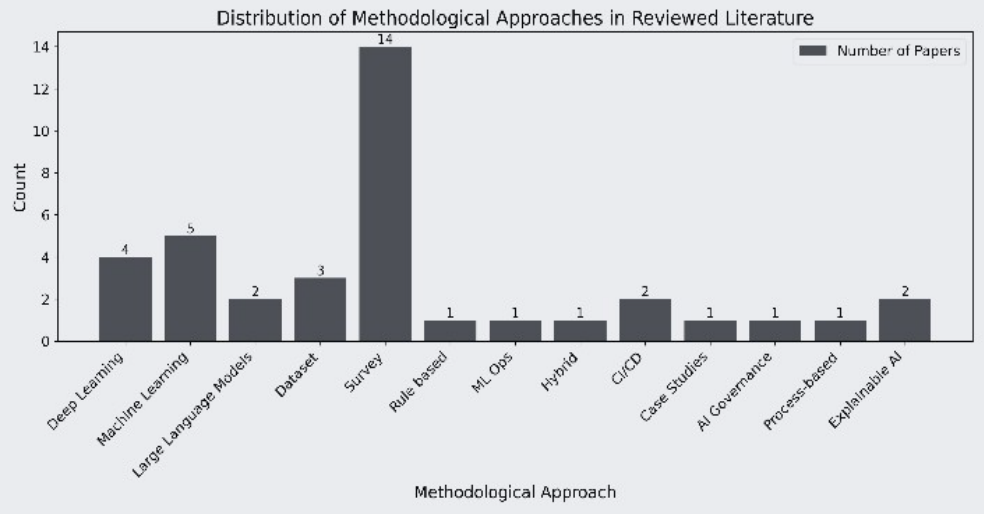


Figure 4. Distribution of levels of analysis in DevSecOps security research

IV. LITERATURE REVIEW

This section critically analyses existing research by grouping studies according to the underlying techniques and their application in DevSecOps environments. Rather than merely describing approaches, the analysis focuses

on practical limitations, comparative strengths, and suitability for explainable, commit-level vulnerability detection.

1. AI Techniques for Vulnerability Detection

Traditional static and dynamic analysis techniques have long been used to detect software vulnerabilities. While these methods are widely adopted due to their ease of integration into development workflows, they suffer from limitations, including high false-positive rates, a lack of contextual understanding, and poor prioritization of security risks. These shortcomings are particularly problematic in CI/CD environments, where frequent alerts can lead to developer fatigue and reduced engagement with security tools [15–17]. sRecent research has explored deep learning and graph-based approaches for vulnerability detection. Models such as VulDeePecker and Devign leverage code semantics, data flow, and program dependencies to identify complex vulnerability patterns [18], [12], [19]. Empirical evaluations show that these models can outperform traditional rule-based techniques in benchmark settings [12], [20]. However, their practical applicability remains limited. Many deep learning models are trained on curated datasets such as DiverseVul, which may not fully capture the noise, diversity, and evolution of real-world codebases. Their black-box nature also limits interpretability, making it difficult for developers to trust or act on predictions [21], [20], [13], [22]. Large language models (LLMs) represent a more recent advancement, enabling contextual reasoning over both code and natural language. These models can analyze complex code structures and provide flexible representations of vulnerabilities. However, they introduce challenges, including hallucinated outputs, unstable reasoning behavior, and high computational overhead, which raise concerns about reliability in security-critical contexts [6], [9], [10], [23], [24], [4], [8]. Overall, AI-based approaches improve detection capability but remain largely model-centric and insufficiently aligned with operational DevSecOps requirements.

2. Commit-Level and Just-in-Time Vulnerability Analysis

Commit-level and just-in-time (JIT) vulnerability detection have emerged as promising approaches for aligning security analysis with CI/CD workflows [3], [8], [22], [14]. Unlike traditional file-level methods, commit-level analysis focuses on incremental code changes, enabling earlier detection of vulnerabilities and improved traceability of security findings to specific modifications. Empirical studies indicate that this approach can reduce the analysis scope and support more efficient remediation, making it well-suited to shift-left security practices [14], [13], [25], [26]. Despite these advantages, current commit-level approaches exhibit several limitations. Many techniques primarily focus on risk prediction rather than providing actionable insights. A significant number of methods also rely on manually engineered features or historical commit data, which can restrict generalization across projects and evolving codebases [19], [26], [25], [24]. Furthermore, commit-level analysis is rarely combined with advanced AI models or explainability mechanisms, resulting in a gap between detection capability and practical usability.

3. Explainable AI in Software Security

Explainable AI (XAI) has become increasingly important in software security, particularly in environments where developers must make rapid and informed decisions. Existing XAI approaches can be broadly categorized into post hoc methods, such as SHAP and LIME, and inherently interpretable models [27], [5], [28]. While these techniques provide insights into model behavior, their practical effectiveness in vulnerability detection remains limited. Post hoc explanations often highlight feature importance or code regions that influence predictions. Still, they may be too abstract or technically complex for developers to interpret in the context of real code changes. Consequently, they may fail to provide actionable remediation guidance. Intrinsically interpretable models offer

improved transparency but can introduce trade-offs involving detection performance or scalability. A critical limitation across existing approaches is the lack of alignment with developer workflows. Most XAI techniques are designed for model analysis rather than integration into CI/CD pipelines. Moreover, explainability is rarely combined with commit-level analysis, limiting its relevance for just-in-time vulnerability detection [20], [23], [29]. This highlights a fundamental gap: current XAI methods improve model transparency but do not adequately support practical decision-making in DevSecOps environments.

4. DevSecOps Security Analytics and Dashboards

DevSecOps platforms commonly employ dashboards to provide visibility into security metrics, including vulnerability counts, severity levels, and compliance status across pipelines. While these tools support monitoring and reporting, they are primarily designed as aggregation interfaces rather than decision-support systems. Most dashboards integrate outputs from multiple security tools without providing insight into how results are generated. As a result, they lack contextual reasoning and fail to explain root causes [24], [22], [30]. This limits their usefulness for developers, who must interpret large volumes of alerts without clear guidance on prioritization or remediation. Current dashboards also rarely adapt to developer workflows or incorporate feedback mechanisms [26], [25], [13], [1].

V. LITERATURE REVIEW

Current DevSecOps environments and vulnerability detection tools differ mainly in their techniques, scope, and practical usefulness [24], [28], [31]. Traditional SAST and DAST tools remain popular because they can integrate into CI/CD pipelines, even though they may generate many false positives and provide limited contextual understanding [32], [22], [16]. Deep learning and graph-based methods can detect complex patterns in code, but they often behave as black boxes and are not consistently evaluated inside real deployment pipelines. Large language model-based approaches improve contextual reasoning; however, their high cost, limited transparency, and limited explainability remain major issues. Commit-level and just-in-time approaches fit CI/CD workflows more naturally, but they often rely on manually designed features and do not clearly explain why a change is risky [5], [28], [24], [17]. Most DevSecOps dashboards collect alerts from different tools and do not effectively support developers in interpreting results [6], [20], [12].

Table 3. Key Quantitative Observations from Reviewed Studies

Aspect	**Observation**
Level of Analysis	More than 50% of studies focus on function-level analysis, while only a small proportion perform commit-level analysis.
Explainability	Approximately 80% of reviewed studies provide no form of explainability for their predictions.
CI/CD Integration	Most approaches are evaluated offline and are not integrated into real CI/CD pipelines.
Dominant Method	Deep learning-based models are prominent in current vulnerability detection research.

XAI Usage	Explainable AI techniques are mostly post-hoc and rarely designed for developer-centric workflows.

Table 4. Comparison of Existing Vulnerability Detection Approaches

Approach	**Summary**
Traditional Security Tools	SAST/DAST; CI/CD compatible, but with frequent false alerts and limited context.
Deep Learning Models	CNN/RNN/GNN; strong pattern detection, but limited explainability and data bias.
LLM-Based Methods	Strong context, but high cost, low explainability, and prediction errors.
Commit-Level / JIT Methods	Good workflow fit, but limited reasoning and reliance on handcrafted features.
Explainable AI Tools	SHAP/LIME provides post hoc explanations but has weak workflow integration.
DevSecOps Dashboards	Good visibility and CI/CD integration, but limited explanation and remediation.

VI. COMPARATIVE ANALYSIS SYNTHESIS COMMIT LEVEL EXPLAINABLE VULNERABILITY

The analysis of the reviewed studies reveals that research on AI-based vulnerability detection, commit-level analysis, explainable AI, and DevSecOps security analytics has largely evolved in isolation rather than as an integrated solution. As a result, while many existing tools are effective at identifying vulnerabilities, they fail to provide developers with actionable and context-aware feedback within CI/CD pipelines [33], [21].

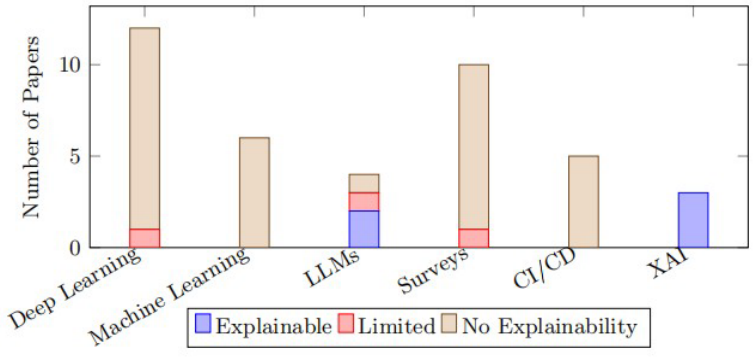


Figure 5. Explainability distribution across different method categories

When these findings are considered collectively, effective DevSecOps security requires the integration of three key dimensions: (i) commit-level change analysis, (ii) AI-based vulnerability prediction, and (iii) explainability tailored to developer workflows. Existing studies rarely address these dimensions simultaneously, leading to fragmented solutions that do not align with real-world development practices. To address this gap, this paper synthesizes a reference framework for commit-level explainable vulnerability intelligence, as illustrated in Figure 6. The framework defines a structured workflow consisting of three main stages: commit-level data processing, AI-based vulnerability prediction, and explainability with developer feedback. Commit-Level Data Processing extracts and preprocesses source-code changes at the commit level to capture fine-grained modifications relevant to security analysis. AI-Based Vulnerability Prediction uses machine learning or deep learning models to estimate the likelihood of vulnerabilities from commit-level features. Explainability and Developer Feedback translate predictions into human-understandable insights, enabling developers to interpret risks and take corrective actions effectively.

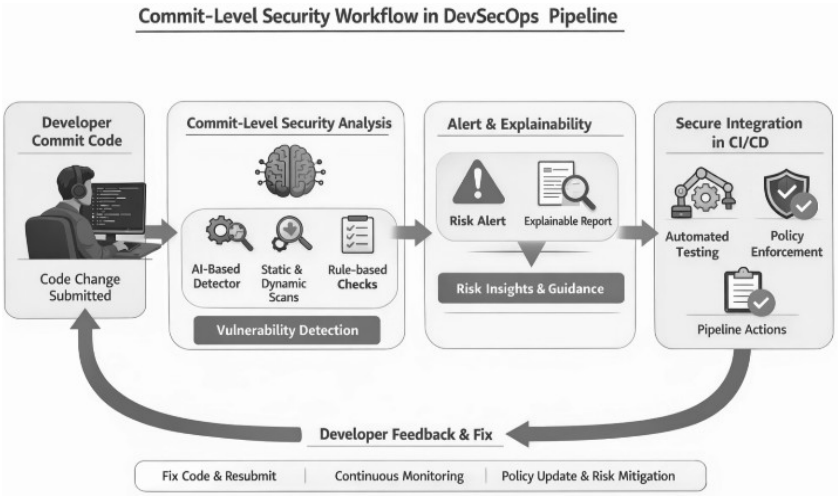


Figure 6. Reference framework for commit-level explainable vulnerability analysis

Unlike approaches that focus on individual components, this framework emphasizes the integration of detection, explanation, and workflow context within CI/CD environments. By aligning vulnerability prediction with commit-level changes and developer-centric explanations, the framework supports faster decision-making, improves trust in automated tools, and enhances the practical usability of AI-driven security systems.

VII. COMPARATIVE ANALYSIS SYNTHESIS COMMIT LEVEL EXPLAINABLE VULNERABILITY

Despite significant advances in AI-based vulnerability detection and DevSecOps practices, several critical challenges continue to limit their practical effectiveness. A fundamental limitation lies in the quality of available datasets. Many widely used vulnerability datasets contain noisy labels, duplicated samples, and weak or indirect links to actual code changes. As a result, models trained on these datasets may fail to generalize effectively, particularly in commit-level analysis scenarios where precise change tracking is essential [19], [22], [24], [34].

Explainability remains another major challenge. Existing XAI techniques frequently provide feature-level or statistical explanations that are difficult for developers to interpret in relation to concrete code modifications or actionable remediation steps. This disconnect reduces developer trust and limits the usability of AI-driven tools

within fast-paced CI/CD environments [28], [13], [24]. The increasing adoption of large language models introduces additional concerns. While these models demonstrate strong contextual understanding, they can produce hallucinated outputs, unstable reasoning, and high computational costs. Such limitations are particularly problematic in security-critical settings, where incorrect or misleading predictions may lead to significant risks [25], [11]. From a systems perspective, current DevSecOps solutions lack effective mechanisms for incorporating developer feedback and operational context into vulnerability analysis. Furthermore, there is no widely accepted architectural model that unifies AI-based detection, explainability, and commit-level analysis within CI/CD pipelines. This fragmentation leads to tools that operate in isolation rather than as part of an integrated security workflow [28], [29], [34], [23]. Existing evaluation methodologies also focus heavily on predictive performance metrics such as accuracy and F1-score, while giving less attention to usability, explanation quality, and real-world pipeline integration. Future research should therefore prioritize high-quality commit-linked datasets, human-centered explainability techniques, standardized architectures, and evaluation strategies that incorporate developer trust, interpretability, and real-world usability.

VIII. CONCLUSION

This paper presented a structured review and synthesis of research on explainable commit-level vulnerability intelligence in DevSecOps pipelines, analyzing 40 selected studies. The review examined AI-based vulnerability detection, just-in-time analysis, explainable AI, and DevSecOps security analytics, providing a classification and critical comparison of existing approaches [33], [21], [9], [2]. The analysis reveals several limitations that hinder the practical adoption of AI-driven security solutions. These include the dominance of black-box models, insufficient explainability, limited use of commit-level analysis, and the absence of standardized integration frameworks. Despite improvements in detection performance, current approaches often fail to provide actionable, context-aware insights that align with real-world development workflows [11], [33], [6]. To address these gaps, this work synthesized a conceptual framework that integrates commit-level analysis, AI-based vulnerability detection, and explainable feedback mechanisms within CI/CD pipelines. The proposed perspective emphasizes aligning technical capabilities with developer needs to support effective and trustworthy security practices. The findings suggest that explainable, commit-level vulnerability intelligence represents a promising direction for both research and practical deployment. Advancing this area will require stronger integration between AI techniques, software engineering processes, and human-centered design principles. This study provides a foundation for future work to develop more reliable, interpretable, and developer-centric security solutions in modern DevSecOps environments.

REFERENCES

[1] Senior DevOps Engineer — Cloud, DevSecOps, and AI/ML, and C. S. Kummarapurugu, “A Framework for Real-Time AI-Driven Secure Code Analysis Integrated with DevSecOps in Cloud-Native CI/CD Pipelines,” IJSREM, vol. 08, no. 11 Nov. 2024, doi: 10.55041/IJSREM17273.

[2] S. Shimmi, H. Okhravi, and M. Rahimi, “AI-Based Software Vulnerability Detection: A Systematic Literature Review,” Jun. 12, 2025, arXiv:2506.10280, doi: 10.48550/arXiv. 2506.10280.

[3] Z. Sheng, Z. Chen, S. Gu, H. Huang, G. Gu, and J. Huang, “LLMs in Software Security: A Survey of Vulnerability Detection Techniques and Insights,” Feb. 12, 2025, arXiv:2502.07049, doi: 10.48550/arXiv. 2502.07049.

[4] G. Bennett, T. Hall, E. Winter, and S. Counsell, “Semgrep*: Improving the Limited Performance of Static Application Security Testing (SAST) Tools,” in Proc. 28th Int. Conf. Evaluation and Assessment in Software Engineering, ACM, 2024, pp. 614–623, doi: 10.1145/3661167.3661262.

[5] "AI-Powered Project Control Dashboards for Proactive Cybersecurity Event Response and Strategic Decision Support in Critical Infrastructure Programs," IJCATR, Aug. 2025, doi: 10.7753/IJCATR1408.1007.

[6] Y. Guo, S. Bettaieb, and F. Casino, "A comprehensive analysis on software vulnerability detection datasets: trends, challenges, and road ahead," Int. J. Inf. Secur., vol. 23, no. 5, pp. 3311–3327, oct. 2024, doi: 10.1007/s10207-024-00888-y.

[7] S. Baker, "AI-Based Threat Detection and Mitigation in DevSecOps."

[8] Q. Mao, Z. Li, X. Hu, K. Liu, X. Xia, and J. Sun, "Towards Explainable Vulnerability Detection with Large Language Models," Sep. 03, 2025, arXiv:2406.09701, doi: 10.48550/arXiv. 2406.09701.

[9] G. Ibba, R. Neykova, M. Ortu, R. Tonelli, S. Counsell, and G. Destefanis, "A machine learning approach to vulnerability detection combining software metrics and topic modeling: Evidence from smart contracts," Machine Learning with Applications, vol. 22, p. 100759, Dec. 2025, doi: 10.1016/j.mlwa.2025.100759.

[10] N. Shiri Harzevili, A. Boaye Belle, J. Wang, S. Wang, Z. M. Jiang, and N. Nagappan, "A Systematic Literature Review on Automated Software Vulnerability Detection Using Machine Learning," ACM Comput. Surv., vol. 57, no. 3, pp. 1–36, Mar. 2025, doi: 10.1145/3699711.

[11] S. Pujar et al., "Analyzing source code vulnerabilities in the D2A dataset with ML ensembles and C-BERT," Empir. Software Eng., vol. 29, no. 2, p. 48, Mar. 2024, doi: 10.1007/s10664-023-10405-9.

[12] A. Papotti, K. Tuma, and F. Massacci, "On the effects of program slicing for vulnerability detection during code inspection," Empir. Software Eng., vol. 30, no. 4, p. 93, Jul. 2025, doi: 10.1007/s10664-025-10636-y.

[13] Y. Ramaswamy, "DevSecOps in Practice: Embedding Security Automation into Agile Software Delivery Pipelines," vol. 31, no. 4, 2023.

[14] S. Reynaud and A. Roxin, "Review of eXplainable artificial intelligence for cybersecurity systems," Discov. Artif. Intell., vol. 5, no. 1, p. 78, May 2025, doi: 10.1007/s44163-025-00318-5.

[15] S. D. Ranwadana, "Vulnerability Management in DevOps CI/CD pipelines."

[16] Z. Li et al., "VulDeePecker: A Deep Learning-Based System for Vulnerability Detection," in Proc. 2018 Network and Distributed System Security Symposium, 2018, doi: 10.14722/ndss. 2018.23158.

[17] J. Wang, "Survey of deep learning models for vulnerability detection," ACE, vol. 92, no. 1, pp. 95–100, Sep. 2024, doi: 10.54254/2755-2721/92/20241722.

[18] V. Nguyen, S. Nepal, X. Yuan, T. Wu, and C. Rudolph, "SAFE: A Novel Approach For Software Vulnerability Detection from Enhancing The Capability of Large Language Models," in Proc. 20th ACM Asia Conf. Computer and Communications Security, ACM, 2025, pp. 392–406, doi: 10.1145/3708821.3736208.

[19] E. Devi, "A Comprehensive Survey on Explainable AI in the Cybersecurity Domain."

[20] Y. Chen, Z. Ding, L. Alowain, X. Chen, and D. Wagner, "DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection," Aug. 09, 2023, arXiv:2304.00409, doi: 10.48550/arXiv.2304.00409.

[21] Y. Zhou, S. Liu, J. Siow, X. Du, and Y. Liu, "Devign: Effective Vulnerability Identification by Learning Comprehensive Program Semantics via Graph Neural Networks," Sep. 08, 2019, arXiv:1909.03496, doi: 10.48550/arXiv. 1909.03496.

[22] F. Binbeshr and M. Imam, "Comparative Analysis of AI-Driven Security Approaches in DevSecOps: Challenges, Solutions, and Future Directions," Apr. 27, 2025, arXiv:2504.19154, doi: 10.48550/arXiv. 2504.19154.

[23] D. Micheal, "AI-Driven Vulnerability Detection and Secure Code Analysis Using Large Language Models."



[24] T. Abdiukov, "Automated security testing in DevSecOps pipelines: Integrating AI-based vulnerability discovery and compliance validation," World J. Adv. Res. Rev., vol. 22, no. 1, pp. 2083–2093, Apr. 2024, doi: 10.30574/wjarr. 2024.22.1.1083.

[25] K. Mohammed, B. Shanmugam, and J. El-Den, "Evolution of DevSecOps and Its Influence on Application Security: A Systematic Literature Review," Technologies, vol. 13, no. 12, p. 548, Nov. 2025, doi: 10.3390/technologies13120548.

[26] X. Zhao, T. Clear, and R. Lal, "Identifying the primary dimensions of DevSecOps: A multi-vocal literature review," Journal of Systems and Software, vol. 214, p. 112063, Aug. 2024, doi: 10.1016/j.jss.2024.112063.

[27] W. R. Nichols, H. Yasar, L. Antunes, and R. McCarthy, "Automated Data for DevSecOps."

[28] N. Alugunuri, "AI for Continuous Security in DevOps (DevSecOps): Integrating Machine Learning into CI/CD Pipelines."

[29] J. Cheenepalli, J. D. Hastings, K. M. Ahmed, and C. Fenner, "Advancing DevSecOps in SMEs: Challenges and Best Practices for Secure CI/CD Pipelines," in 2025 13th International Symposium on Digital Forensics and Security (ISDFS), Apr. 2025, pp. 1–6, doi: 10.1109/ISDFS65363.2025.11011960.

[30] B. Casey, J. C. S. Santos, and G. Perry, "A Survey of Source Code Representations for Machine Learning-Based Cybersecurity Tasks," ACM Comput. Surv., vol. 57, no. 8, pp. 1–41, Aug. 2025, doi: 10.1145/3721977.

[31] A. Challande, R. David, and G. Renault, "Building a Commit-level Dataset of Real-world Vulnerabilities," in Proc. 12th ACM Conf. Data and Application Security and Privacy, ACM, 2022, pp. 101–106, doi: 10.1145/3508398.3511495.

[32] R. Croft, M. A. Babar, and M. Kholoosi, "Data Quality for Software Vulnerability Datasets," Jan. 13, 2023, arXiv:2301.05456, doi: 10.48550/arXiv. 2301.05456.

[33] S. Nguyen, T.-T. Nguyen, T. T. Vu, T.-D. Do, K.-T. Ngo, and H. D. Vo, "Code-centric Learning-based Just-In-Time Vulnerability Detection," Apr. 17, 2023, arXiv:2304.08396, doi: 10.48550/arXiv.2304.0839

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

The Illusion of Cultural Competence: Evaluating ChatGPT-Generated Kazakh Proverbs

Shynar Auyelbekova
Center for Multidisciplinary Education
SDU University
Almaty, Kazakhstan
shynar.auyelbekova@sdu.edu.kz

Gulbanu Mukhiyayeva
Independent Researcher
Almaty, Kazakhstan
banumuhya@gmail.com

Abstract
As Large Language Models (LLMs) like ChatGPT become global information utilities, their performance in low–resource languages remains a critical area of inquiry. While these models demonstrate syntactic fluency in languages such as Kazakh, their ability to reproduce authentic cultural knowledge–specifically fixed paremiological units (proverbs)–is under–evaluated. This study assesses the validity of LLM–generated Kazakh proverbs using a binary expert-verification protocol administered by two experts in Turkic philology. We generated 1,500 outputs across three thematic domains–Labor, Humanity, and Knowledge–and classified each against authoritative lexicographical sources as Authentic or Fabricated. Results reveal a critical dissonance: although outputs were generally grammatically well–formed, only 8.9% of unique strings were authentic Kazakh proverbs. Authenticity correlated strongly with generation frequency ($r = 0.87$), indicating that the model relies on a small memorized core of high–frequency proverbs while hallucinating plausible–sounding fabrications in the long tail. These findings quantify the "illusion of cultural competence" in current LLMs and provide a reproducible methodology for evaluating cultural knowledge generation in low–resource languages.

Keyword: generative AI, low–resource languages, Kazakh language, cultural hallucination, paremiology, LLM evaluation

I. INTRODUCTION

Since its public release, the ChatGPT large language model (LLM) has been utilized by hundreds of millions of active users [1]. As of November 2025, the system has approximately 800 million weekly active users globally and continues to handle over 2.5 billion daily requests [2]. ChatGPT is also widely used in Kazakhstan [3]. Given this widespread adoption, it is critical to assess the nature of ChatGPT's generative outputs. However, systematic evaluation of LLM performance on low–resource languages–particularly for cultural knowledge generation–remains limited, creating gaps in our understanding of model capabilities and limitations beyond high–resource language benchmarks [4–7]. Proverbs represent a challenging test case for evaluating cultural knowledge in NLP systems. Unlike generic text generation, proverb production requires both linguistic competence (grammatical correctness, stylistic appropriateness) and cultural authenticity (correspondence to documented traditional knowledge) [8]. For Kazakh, a Turkic language with approximately 15 million speakers but a limited digital corpus representation [6], this evaluation is particularly relevant as speakers increasingly rely on LLMs for language–related tasks. Existing NLP research on low–resource languagses primarily focuses on machine translation or on standard benchmarks [4], [9]. Studies of LLM-generated cultural content predominantly focus

on high–resource languages [10] or examine translation rather than generation [7], [11], [12]. This leaves critical questions unaddressed, which we state as the three research questions guiding this study: (RQ1) Can LLMs authentically generate cultural artifacts in low–resource languages? (RQ2) How do these systems balance memorized traditional knowledge against generative pattern completion? (RQ3) What evaluation methodologies can distinguish linguistic competence from cultural authenticity? This study provides the first systematic computational and linguistic evaluation of LLM–generated Kazakh proverbs across three culturally significant domains: Labor (еңбек), Humanity/Ethics (адамгершілік), and Knowledge (білім). Through iterative generation and expert verification by two specialists in Turkic philology, we document patterns of linguistic accuracy, cultural authenticity, and the relationship between generation frequency and knowledge quality. In doing so, we contribute a reproducible expert–verified protocol, evidence separating grammatical well–formedness from cultural authenticity, and a frequency–based diagnostic distinguishing memorized retrieval from pattern completion.

II. RELATED WORK

1. LLM Evaluation on Low–Resource Languages

Recent multilingual evaluations document persistent performance disparities between high– and low–resource languages. Ahuja et al.'s [9] MEGA benchmark evaluated GPT–3.5, GPT -4, and BLOOMZ across 70 typologically diverse languages on 16 NLP datasets. They found significant performance disparities between English and non–English languages, particularly for low–resource languages with non–Latin scripts, where tokenizer quality (measured by sub–word fertility) negatively correlated with performance and fine–tuned models often outperformed generative LLMs. Bang et al. [13] extended these findings in a comprehensive multitask evaluation of ChatGPT, noting a critical performance asymmetry: while ChatGPT demonstrates competence in understanding low–resource and non–Latin script languages, its ability to generate text in those languages is significantly weaker. Joshi et al. [4] developed a taxonomy that classifies the world's languages into six tiers based on the availability of digital resources, highlighting a stark disparity: a small number of "Winner" languages receive the vast majority of NLP research and data investment. They argue this creates a "typological echo–chamber," leaving languages with moderate to low digital infrastructure–a category that includes Kazakh–substantially underrepresented in NLP research. While their work focuses on resource distribution, it underscores the need for evaluations that go beyond standard NLP benchmarks to examine more nuanced capabilities, such as generating cultural knowledge in these low–resource languages.

2. Cultural Knowledge in Language Models

Theoretical work on LLM knowledge representation raises fundamental questions about the nature of cultural understanding in neural models. Bender et al. [5] characterize these systems as "stochastic parrots"–models that haphazardly stitch together sequences of linguistic forms based on probabilistic patterns observed in vast training data, but without any underlying reference to meaning. They argue that while models achieve impressive surface fluency, they lack genuine semantic comprehension because their training is restricted to linguistic form, denying them access to the communicative intent and shared world models that underpin human language [5]. This disconnect is particularly acute for cultural artifacts such as proverbs; Bender et al. posit that any perceived coherence in these outputs is often an "illusion" created by human listeners who are predisposed to project meaning and intent onto fluent strings, even when the generating system lacks conceptual grounding [5]. Navigli et al. [14] documented systematic biases in LLM cultural representations, noting that training data is heavily English–centric and over–represents high–resource linguistic communities, leading to models that reflect the values of dominant cultures while marginalizing others. The challenge of representing non–Western cultural

heritage in LLMs has recently been highlighted by the Jawaher benchmark as well [7]. Magdy et al. evaluated LLMs across 20 Arabic dialects and found that while models demonstrate high linguistic surface–level performance, they suffer from a significant 'cultural gap.' Their study concludes that current models remain biased toward Anglo–centric perspectives, frequently failing to capture the historical and contextual depth of traditional proverbs–a phenomenon we investigate here within the Kazakh linguistic context.

3. Proverb Generation as a Benchmark

Beyond translation, proverbs have been established as a challenging benchmark for evaluating abstract and analogical reasoning in LLMs. Ghosh and Srivastava [10] introduced the ePiC dataset to demonstrate that even when models achieve high performance on standard NLP tasks, they often struggle to move beyond surface–level lexical overlaps to achieve a true abstract understanding of proverbial wisdom. Recent scholarship has increasingly turned to proverbs as a rigorous metric for cultural alignment in LLMs. Magdy et al. [7] (Arabic), Azime et al. [11] (Ethiopian languages), and Almeida et al. [12] (Portuguese) have all demonstrated that models trained primarily on English data fail to capture the specific regional nuances of proverb usage. While these studies largely utilized multiple–choice or retrieval tasks, our work extends this line of inquiry into the generative domain for Kazakh, exposing the model's tendency to hallucinate plausible–sounding but culturally invalid artifacts when forced to generate without constraints. Evaluating generation is crucial because it mimics real–world usage: users often ask LLMs to "write a speech using Kazakh proverbs" or "explain Kazakh culture" [3]. If the generated content is fabricated, the model actively contributes to cultural erosion [5].

III. METHODOLOGY

1. Data Generation and Collection

To ensure a robust and statistically significant sample size, this study employed an iterative prompting strategy designed to capture both the model's consistency and its generative variability. Data was collected in late 2024 using OpenAI's GPT–3.5 Turbo at its default generation settings, reflecting the configuration most users encounter rather than a tuned research configuration. The exact-date snapshot was not recorded and is no longer available (Section V.5). We used a standardized, zero–shot instruction written in Kazakh: "Generate 10 authentic Kazakh proverbs about [TOPIC] in the Kazakh language." To test the model's ability to generalize across different conceptual depths, we selected three distinct thematic domains: Labor (Еңбек), representing concrete, experience–based concepts; Humanity and Ethics (Адамгершілік), representing abstract moral philosophy; and Knowledge (Білім), representing metaphorical and educational themes. For each topic, the prompt was executed 50 times using fresh context windows to eliminate history bias and prevent the model from referencing its own previous outputs. A single zero–shot prompt was used because it reflects how non–specialist users typically query the model; the effect of prompt variation is addressed in Section V.5. This procedure resulted in a raw corpus of 1,500 proverb outputs.

2. Verification Procedure and Binary Classification

The primary methodological contribution of this study is the shift from automated metrics toward an expert–led verification process. Initially, raw text was normalized by removing punctuation and case differences to facilitate deduplication and identify unique strings. Following the human evaluation framework established in recent paremiological benchmarks [7], we engaged two qualified experts in Turkic philology to ensure that the generated strings were not only grammatically correct but also culturally legitimate. The first expert assigned an initial binary label (Authentic or Fabricated) to each unique output by cross–referencing it against the

authoritative paremiological collections described below. The second expert independently reviewed all classifications. Cases of disagreement were resolved through joint discussion against the source corpora before producing the final dataset. This two–reviewer procedure reduces dependence on any single expert's judgment. To ensure rigorous binary verification of authenticity, we used a consolidated ground–truth corpus comprising four authoritative paremiological collections: the comprehensive Khalyq danalyghy [15] and Babalar sözi [16], as well as specialized collections compiled by Turmanzhanov [17] and Asqar [18]. We operationalized authenticity through a strict binary coding system. An output was coded as "1" (Authentic) if it was documented in the source material, including acceptable oral variations or minor synonym swaps that preserved the proverb's lexical core, syntactic structure, and semantic content. An output was coded as "0" (Fabricated) if it was absent from authoritative sources; this category included plausible AI–invented phrases, literal translations (calques) of English or Russian idioms, and semantic nonsense. An output was only marked as Authentic if it was documented in at least one of these recognized scholarly sources.

3. Acceptance of Variants and Sensitivity Analysis

Because Kazakh proverbs circulate as oral tradition with documented regional variation, we explicitly defined three levels of acceptable variation. Level 1 (lexical): minor synonym swaps that do not alter meaning (e.g., Еңбек етсең vs. Еңбек қылсаң, both meaning "if you work"). Level 2 (morphosyntactic): aspect or tense alternations that preserve propositional content. Level 3 (structural): partial reorderings that retain the parallel–couplet structure. Outputs falling within these levels were classified as Authentic if a matching base form existed in at least one source. Outputs that introduced new lexical content, blended distinct proverbs, or replaced culturally specific tokens with foreign idiomatic content were classified as Fabricated. To estimate possible false negatives arising from the finite scope of the reference corpora, we conducted a sensitivity analysis. We re–examined all items classified as Fabricated for documented partial overlap with the reference corpora, identifying 26 unique strings that exhibited a recognizable but non–canonical relationship to attested proverbs (e.g., partial token overlap or thematic echo without full structural match). Reclassifying these items as plausibly Authentic raises the unique authenticity rate from 8.9% to 12.3%–a range of ±3.4 percentage points. This bound demonstrates that the headline finding (approximately 90% fabrication) is robust to plausible variation in coding strictness.

4. Evaluation Metrics

To quantify the model's performance, we established two primary metrics. The Unique Authenticity Rate measures the percentage of distinct, unique strings that constitute genuine traditional proverbs. The Weighted Authenticity Rate accounts for repetition frequency by measuring the percentage of the 1,500 generated outputs that are authentic. These dual metrics allow us to distinguish between the model's total knowledge base and its reliability in a typical user interaction.

IV. ANALYSIS AND RESULTS

1. Quantitative Overview of Generated Data

Across the 1,500 total generation requests, ChatGPT produced 767 unique strings, indicating a significant repetition rate of approximately 48.9%. As summarized in Table 1, the data reveal a stark generative tendency that undermines cultural accuracy. Out of the 767 unique outputs, expert verification confirmed that only 68 items (8.9%) represented authentic traditional proverbs. Notably, the Knowledge (Білім) domain had the lowest unique authenticity rate, at 7.8%, compared with Labor (9.3%) and Humanity (9.2%). This indicates that the

abstract and metaphorical depth of educational proverbs poses a particular challenge for the model. Overall, approximately 91% of the model's unique "wise sayings" are fabrications, highlighting a severe reliability gap when the model is tasked with retrieving specific cultural heritage artifacts.

Table 1. Performance Metrics by Domain

Metric	Labor	Humanity	Knowledge	Total
Total Generated	500	500	500	1,500
Unique Proverbs	259	303	205	767
Repetition Rate	48.2%	39.4%	59.0%	48.9%
Unique Authenticity	9.3%	9.2%	7.8%	8.9%
Weighted Authenticity	43.4%	29.8%	47.6%	40.3%

Note. Repetition = (total − unique) / total. Unique authenticity = verified / unique. Weighted authenticity = Σ(authentic frequencies) / Σ(all frequencies). All verified against authoritative sources [15–18].

2. The Frequency–Authenticity Correlation

A pivotal finding of this analysis is the significant relationship between generation frequency and factual validity. To examine this, we categorized unique strings into three tiers: High Frequency (appearing 10 or more times), Medium Frequency (3–9 times), and Low Frequency (1–2 times), the latter representing the "long tail" of AI generation. The results reveal that High–Frequency outputs achieved an authenticity rate of 88.0%. In contrast, the authenticity rate dropped to 53.3% for Medium–Frequency items and plummeted to a negligible 3.2% for Low–Frequency items (Figure 1).

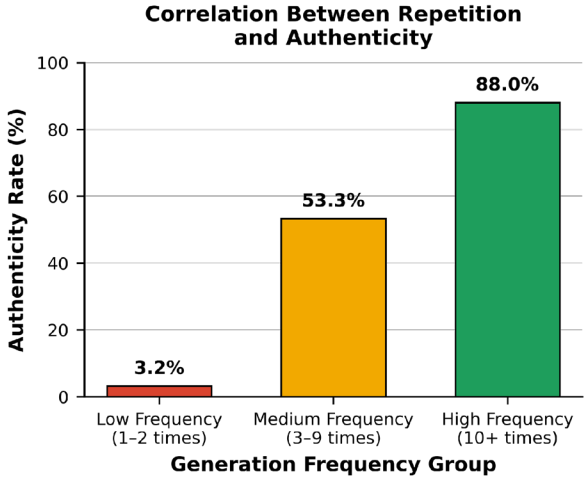


Figure 1. Correlation between generation frequency and cultural authenticity. The sharp decline in authenticity for low–frequency outputs suggests a transition from memorized retrieval to stochastic fabrication. To confirm the robustness of this pattern, we computed a Pearson correlation between generation frequency and authenticity across all unique outputs, yielding $r = 0.87$ ($p < 0.001$). This strong positive correlation indicates that the relationship between repetition and authenticity is statistically significant rather than coincidental.

3. Cross–Domain Analysis

Performance consistency across domains was notable. Unique authenticity varied within a narrow range of 1.5 percentage points: Labor 9.3%, Humanity 9.2%, Knowledge 7.8%. However, weighted authenticity showed greater variation: Labor 43.4%, Humanity 29.8%, Knowledge 47.6%. The Knowledge domain pattern–lowest unique authenticity but highest weighted authenticity–reveals an important asymmetry: the model retrieves a very small set of well–memorized Knowledge proverbs (e.g., the famous Оқу инемен құдық қазғандай) with very high repetition, but exhibits the weakest generative breadth in this conceptually abstract domain. By contrast, Labor proverbs show broader memorized coverage but lower per–item repetition. This pattern suggests similar generation mechanisms across domains, but differential knowledge depth tied to the conceptual character of each domain.

4. Qualitative Analysis of Fabricated Outputs

While the quantitative data show a high rate of failure, a qualitative review of the Fabricated category identifies specific failure modes that illuminate how the model attempts to imitate Kazakh proverbial form without access to authentic cultural content. From a paremiological perspective, the fabrications fall into four recognizable categories, each pointing to a different limitation of the model. Semantic Incoherence (Nonsense). The model occasionally produces sentences that violate basic semantic coherence by blending tokens from unrelated proverbs. The output “Еңбек етсең елік жылан да жығылады” (If you work, even the roe deer snake falls) appears to be a hallucinated blend of a labor–themed conditional construction and the famous proverb on the warmth of speech (Жылы–жылы сөйлесең, жылан інінен шығады). The insertion of елік (roe deer) before жылан (snake) produces a semantically impossible compound (“roe–deer–snake”) that no Kazakh speaker would accept. Such outputs are useful diagnostically: they reveal that the model is attempting to assemble proverbial–sounding strings from co–occurring tokens rather than reproducing memorized cultural units.

Cultural Calques. A second failure mode is the direct translation of foreign idioms into Kazakh grammatical form. The output “Балық берме, балық аулауды үйрет” is a word–for–word calque of the English saying “Give a man a fish…; teach him to fish…” While grammatically Kazakh, it carries no resonance in the Kazakh oral tradition. This category is particularly concerning for cultural preservation, because such outputs could plausibly be mistaken for indigenous wisdom by users unfamiliar with the original foreign idiom—distorted Variants of Real Proverbs. The most linguistically revealing category consists of outputs that closely resemble authentic proverbs but with one or two altered lexical items, yielding variants attested by no source. For example, the fabricated “Еңбек қылсаң ерінбей, пайда болар әр ісің” (If you work without laziness, every task will bring benefit) is a variation of the attested “Еңбек қылсаң ерінбей, тояды қарның тіленбей” (If you work without laziness, your stomach will be full without begging). The first half of the couplet is preserved exactly, but the second half is replaced with a generic abstraction. The authentic version uses the concrete image of a full stomach (тояды қарның); the fabrication substitutes a vague phrase about “benefit” (пайда). Similar near–misses appear across the Labor domain (e.g., the attested “Еңбекшінің наны тәтті, жалқаудың жаны тәтті” becoming the fabricated “Жалқаудың жаны тәтті, еңбекшінің күні тәтті”). The frequency of such micro–distortions suggests that the model has partial access to canonical Kazakh proverbs but lacks the lexical specificity to reproduce them precisely, defaulting to semantically adjacent but historically unattested substitutions plausible but Non–existent Wisdom. The most challenging category for evaluation consists of outputs that exhibit the full structural apparatus of Kazakh proverbs–parallel–couplet construction, vowel harmony, end–rhyme, and conventional thematic content–yet are absent from any documented source. The output “Еңбек етпегеннің несібесі аз” (He who does not work has little share) is grammatically and stylistically indistinguishable from attested labor proverbs; it would not arouse suspicion in a casual reader. These outputs reveal that the model has

internalized the form of Kazakh paremiological discourse without internalizing the specific content of the oral tradition, and they represent the principal risk to applications in cultural transmission. These four categories also exhibit a notable distributional pattern. Semantic Incoherence and Cultural Calques are easily identifiable by fluent speakers and would not survive casual scrutiny. Distorted Variants and Plausible but Non–existent Wisdom, however, can deceive even attentive readers without access to authoritative reference works, and these two categories together account for the majority of fabricated outputs in our sample. From a cultural preservation standpoint, the latter two categories warrant the greatest concern.

Table 2.Representative Examples of Verified vs. Fabricated Outputs

Category	**ChatGPT Output (Kazakh)**	**English Translation / Analysis**	**Expert Verdict**
Knowledge	Оқу инемен құдық қазғандай.	“Learning is like digging a well with a needle.”	Authentic
Labor	Еңбек етсең елік жылан да жығылады.	“If you work, even a roe deer snake falls.” (Logical impossibility.)	Fabricated (Nonsense)
Ethics	Балық берме, балық аулауды үйрет.	“Give a man a fish, teach him to fish.” (Direct English translation.)	Fabricated (Calque)
Labor	Жатқан жыланға да тау келеді.	“The mountain comes to the lying snake.” (Merged metaphors.)	Fabricated (Nonsense)
Labor	Еңбек етпегеннің несібесі аз.	“He who does not work has little share.” (Plausible but unrecorded.)	Fabricated (Plausible)

V. DISCUSSION

1. The “Illusion of Fluency”

The most concerning finding for the NLP community is the disconnect between linguistic fluency and cultural accuracy. A user interacting with ChatGPT would see correctly spelled, grammatically complex Kazakh sentences. Without expert philological knowledge, a user could plausibly accept a substantial portion of these fabrications as real–particularly those in the “Plausible but Non–existent Wisdom” category, which exhibit the structural features of authentic proverbs (rhyme, vowel harmony, parallel–couplet construction). Even when more obvious failures (semantic nonsense or foreign calques) are easily discernible to fluent speakers, the volume of grammatically and stylistically authentic–sounding fabrications poses a risk to educational and cultural transmission applications (Fig. 2), where AI might inadvertently rewrite local traditions with synthetic wisdom.

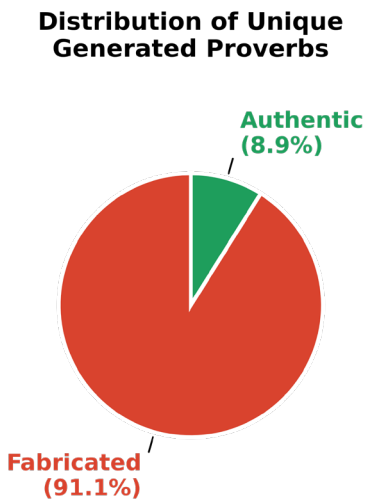


Figure 2. Distribution of authenticity among uniquely generated proverbs. Despite general linguistic well–formedness, the vast majority of unique outputs constitute cultural hallucinations.

2. Memorization vs. Pattern Completion

The frequency–authenticity correlation suggests that ChatGPT's outputs reflect two distinct underlying behaviors. For high–frequency outputs–the top ≈25 most famous Kazakh proverbs (e.g., Оқу инемен құдық қазғандай)–the model functions as a retrieval system. It has seen these specific token sequences many thousands of times in training data, and reproduces them reliably. For low–frequency outputs, the model has limited or no memorized referent and instead generates by pattern completion: predicting what a proverb should sound like based on probabilistic associations between the topic and related tokens (e.g., "Labor" with "Hardship" and "Bread"), producing grammatically well–formed but culturally unattested sentences.

3. Possible Technical Explanations

Two technical factors likely contribute to this pattern. First, frequent proverbs appear repeatedly across publicly available web sources–textbooks, blog posts, news articles, and Wikipedia entries–while less prominent proverbs appear only in specialized lexicographical works, plausibly affecting the model's exposure to each item during training. Second, prior NLP research notes that Kazakh's Cyrillic script is tokenized inefficiently by models trained primarily on English [9], which may further degrade performance on multiword expressions such as proverbs. A full technical investigation of these factors is beyond the scope of the present linguistic study.

4. Practical Implications

For developers building tools for low–resource languages, this study suggests that temperature settings and generation limits must be carefully tuned. Relying on an LLM's "creativity" for cultural tasks is detrimental. Instead, approaches such as Retrieval–Augmented Generation (RAG), which force the model to look up data in a trusted corpus (such as Babalar sözi) before generating, are essential for cultural reliability. Our results also suggest a practical heuristic for end users: in low–resource cultural–knowledge generation, repetition across

independent prompt instances is a reasonable proxy for authenticity. Outputs that the model produces only once should be treated with suspicion.

5. Limitations and Future Work

Several limitations should be acknowledged. First, our verification relies on published lexicographical sources, which, despite their authority, do not capture the full oral tradition; rare regional variants may have been undercounted, although the sensitivity analysis in Section III.3 bounds the resulting uncertainty to approximately ±3.4 percentage points. Second, the prompt–elicitation procedure used a single zero–shot Kazakh language instruction; while this reflects naturalistic user interaction, the model's behavior under source–constrained prompts (e.g., explicitly citing Babalar sözi) or chain–of–thought prompts may differ. We did not test multiple prompt variants because the original GPT–OpenAI has since deprecated 3.5 Turbo, the snapshot used for data collection, and it is no longer accessible for replication; running new prompts against the current model would conflate prompt effects with model–version effects. Third, our study evaluates a single model version. The memorization–versus–generation pattern documented here is an interpretation of LLM behavior on low–resource cultural knowledge, not a claim about all LLMs; replication on newer models (GPT–4, Claude, Llama–family models) and on additional low–resource Turkic languages (e.g., Kyrgyz, Uzbek, Tatar) would clarify which findings reflect general LLM behavior and which are specific to GPT–3.5 Turbo on Kazakh. Fourth, while domain selection covered three culturally significant areas, additional domains (e.g., agriculture, family, hospitality) could reveal further variation in the depth of memorization. Future work should pursue three directions. First, comparative evaluation across newer model generations and multiple Turkic languages would test the cross–lingual generality of this memorization–versus–generation pattern and isolate effects attributable to tokenizer geometry, training data scarcity, and model scale. Second, prompt–engineering and fine–tuning interventions–particularly retrieval–augmented architectures using curated folkloric corpora such as Babalar sözi–should be evaluated as remedies for the long–tail collapse. Third, following Almeida et al. [12], future studies could pair Kazakh proverbs with short narrative contexts to evaluate semantic grounding rather than mere string retrieval. This direction is particularly important given our finding that ChatGPT fails at the simpler zero–shot generation task: a system that cannot reliably produce attested proverbs is unlikely to apply them appropriately in context, and explicit contextual–grounding benchmarks are needed to determine whether next–generation models suffer from the same surface–fluency–without–meaning gap observed here.

VI. CONCLUSION

This study provides a systematic evaluation of LLM-generated cultural knowledge for a low–resource language, documenting a fundamental asymmetry between linguistic competence and cultural authenticity. While achieving high grammatical correctness, ChatGPT produces predominantly inauthentic content (91.1% fabrication rate at the level of unique strings; sensitivity range 87.7–91.1%), relying on a small core of memorized proverbs supplemented by pattern–generated alternatives. The strong frequency–authenticity correlation ($r = 0.87$, $p < 0.001$) supports a memorization–versus–pattern–completion account with implications for NLP evaluation, low–resource language applications, and responsible AI deployment in cultural contexts. As LLMs achieve greater linguistic sophistication and global deployment, understanding the limitations of cultural knowledge representation becomes critical for equitable access and responsible use. This work provides an empirical foundation and reproducible methodology for evaluating cultural authenticity beyond surface–level linguistic competence–essential for developing AI systems that serve diverse linguistic communities effectively and appropriately.

REFERENCES

[1]. [1] OpenAI, “AI as a healthcare ally: How Americans are navigating the system with ChatGPT,” OpenAI Report, Jan. 2026. [Online]. Available: https://cdn.openai.com/pdf/2cb29276–68cd–4ec6–a5f4–c01c5e7a36e9/OpenAI–AI–as–a–Healthcare–Ally–Jan–2026.pdf Accessed: May 15, 2026.

[2]. [2] E. Roth, “OpenAI says ChatGPT is now processing 2 billion prompts a day,” The Verge, Jul. 21, 2025. [Online]. Available: https://www.theverge.com/news/710867/openai–chatgpt–daily–prompts–2–billion Accessed: May 15, 2026.

[3]. [3] Kazakhstan Institute for Strategic Studies (KazISS), “Kazakhstanis most often turn to digital tools for study and self–development: KazISS survey,” Nov. 24, 2025. [Online]. Available: https://kisi.kz/en/kazakhstanis–most–often–turn–to–digital–tools–for–study–and–self–development–kaziss–survey/ Accessed: May 15, 2026.

[4]. [4] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, “The state and fate of linguistic diversity in NLP,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguist, 2020, pp. 6282–6293.

[5]. [5] E. M. Bender, T. Gebru, A. McMillan–Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in Proc. Conf. Fairness, Accountability, and Transparency (FAccT), 2021, pp. 610–623.

[6]. [6] M. Togmanov et al., “KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan,” in Proc. 63rd Annu. Meeting Assoc. Comput. Linguist. (Vol. 1), 2025, pp. 14403–14416.

[7]. [7] S. M. Magdy et al., “Jawaher: A multidialectal dataset of Arabic proverbs for LLM benchmarking,” in Proc. 2025 Conf. Nations of the Americas Chapter Assoc. Comput. Linguist. (NAACL), 2025, pp. 12320–12341.

[8]. [8] W. Mieder, Proverbs: A Handbook. Westport, CT: Greenwood Press, 2004.

[9]. [9] K. Ahuja et al., “MEGA: Multilingual evaluation of generative AI,” in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), 2023, pp. 4232–4267.

[10]. [10] S. Ghosh and S. Srivastava, “ePiC: Employing proverbs in context as a benchmark for abstract language understanding,” in Proc. 60th Annu. Meeting Assoc. Comput. Linguist. (Vol. 1: Long Papers), 2022, pp. 3214–3234.

[11]. [11] I. A. Azime et al., “ProverbEval: Exploring LLM evaluation challenges for low–resource language understanding,” in Findings Assoc. Comput. Linguist.: NAACL 2025, 2025, pp. 6265–6281.

[12]. [12] T. S. Almeida, G. K. Bonás, and J. G. A. Santos, “BRoverbs – Measuring how much LLMs understand Portuguese proverbs,” J. Braz. Comput. Soc., vol. 31, no. 1, pp. 1077–1087, 2025.

[13]. [13] Y. Bang et al., “A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity,” in Proc. 13th Int. Joint Conf. Nat. Lang. Process. 3rd Conf. Asia–Pacific Chapter Assoc. Comput. Linguist, 2023, pp. 675–718.

[14]. [14] R. Navigli, S. Conia, and B. Ross, “Biases in large language models: Origins, inventory, and discussion,” ACM J. Data Inform. Quality, vol. 15, no. 2, pp. 1–21, 2023.

[15]. [15] A. T. Kaydar, Khalyq danalyghy (qazaq maqal–mätelderıniñ tüsındırme sözdıgı jäne zerttey) [Folk Wisdom: Explanatory Dictionary and Study of Kazakh Proverbs and Sayings]. Almaty, Kazakhstan: Toghanay T, 2004.

[16]. [16] K. Alpysbaeva, T. Alibekov, and S. Kaskabasov, Eds., Babalar sözi: Jüz tomdyq [Words of the Ancestors: 100 Volumes], vol. 65, Qazaq maqal–mätelderı [Kazakh Proverbs and Sayings]. Astana, Kazakhstan: Foliant, 2010.

[17]. [17] Ö. Turmanzhanov and K. Akhmetbekuly, Comps., Qazaqtyñ maqal–mätelderı [Kazakh Proverbs]. Almaty, Kazakhstan: Raritet, 2004.

[18]. [18] O. Asqar, Comp., Qazaq jäne älem halyqtarynyñ tañdamaly maqal–mätelderı [Selected Proverbs of Kazakh and World Nations]. Almaty, Kazakhstan: Balausa, 2005.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A QUBO-Based Reformulation and Quantum Solution Framework for Nonlinear Constrained Optimization Problems

Cumali Yaşar
Department of Computer Education and Instructional Technology
Çanakkale Onsekiz Mart University
Çanakkale, Türkiye
cyasar@comu.edu.tr
ORCID: 0000-0002-0065-3752

Yerlan Izbassarov
Department of Software Engineering
Khalel Dosmukhamedov Atyrau University
Atyrau, Kazakhstan
erlan_1081@mail.ru
ORCID: 0000-0002-1917-2086

Abstract
Nonlinear constrained optimization problems arise in many domains, including engineering design, production planning, machine learning, and decision-support systems. Although classical optimization algorithms can effectively solve many small- and medium-scale instances, reformulating such problems for quantum optimization remains a challenging modeling task. This study proposes a QUBO-based reformulation framework for transforming nonlinear constrained optimization problems into a binary quadratic form suitable for quantum-inspired and quantum-based solution approaches. The proposed framework includes variable discretization, binary encoding, penalty-based constraint embedding, and QUBO matrix construction. A nonlinear resource-constrained production planning problem is used as a case study. The problem is first solved using a classical constrained optimizer, then reformulated as a QUBO model and solved using classical QUBO search and a quantum-oriented optimization approach. The results are evaluated in terms of objective value, constraint violation, solution feasibility, and computational behavior. The study provides a reproducible Python-based workflow and establishes a foundation for extending nonlinear constrained optimization problems toward quantum programming environments. The proposed approach helps bridge classical nonlinear optimization and emerging quantum optimization methods.

Keywords: nonlinear constrained optimization, QUBO, quantum optimization, quantum programming, QAOA, binary encoding, penalty method.

I. INTRODUCTION

Nonlinear constrained optimization problems arise in many domains, including engineering design, production planning, resource allocation, logistics, finance, machine learning, and decision-support systems. These problems require optimizing a nonlinear objective function while satisfying equality or inequality constraints. Although classical methods such as sequential quadratic programming, interior-point methods, penalty methods, and gradient-based algorithms can solve many instances effectively, their performance may be affected by nonconvexity, local optima, constraint complexity, and the dimensionality of the search space. Quantum computing and quantum-inspired optimization have recently attracted attention as alternative computational paradigms for complex optimization problems. In this context, Quadratic Unconstrained Binary Optimization, known as QUBO, provides a compact bridge between classical optimization models and quantum-oriented algorithms. Many quantum optimization approaches, including quantum annealing and the Quantum Approximate Optimization Algorithm, require problems to be represented in binary quadratic form. Therefore, transforming nonlinear constrained optimization problems into QUBO form is a critical step. However, this transformation is not straightforward. Nonlinear constrained optimization problems often include continuous decision variables, nonlinear objective terms, and multiple constraints. To make such problems suitable for



QUBO-based solution methods, continuous variables must be discretized and encoded as binary variables, and constraints must be embedded in the objective function via penalty terms. The quality of this transformation directly affects feasibility, QUBO matrix size, and computational behavior. This study proposes a QUBO-based reformulation framework for nonlinear constrained optimization problems. A nonlinear resource-constrained production planning problem is used as a case study. The problem is first solved using a classical constrained optimizer, then transformed into QUBO form via variable scaling, binary encoding, slack-variable representation, and penalty-based constraint embedding. The resulting QUBO model is evaluated using exact enumeration, random search, simulated annealing, and a PennyLane-based QAOA demonstration. The main contributions of this study are threefold. First, a transparent QUBO reformulation workflow is presented for transforming a nonlinear constrained optimization problem into a binary quadratic form. Second, the proposed transformation is validated on a resource-constrained production planning problem by comparing the classical SLSQP solution with the exact QUBO solution. Third, a reproducible Python-based workflow is developed to connect classical optimization, QUBO modeling, and quantum-oriented optimization. This paper does not claim a quantum advantage. Instead, it focuses on the modeling pathway from classical nonlinear constrained optimization to QUBO-based quantum-oriented optimization. The remainder of the paper is organized as follows. Section II reviews related work. Section III presents the problem formulation. Section IV explains the QUBO reformulation framework. Section V describes the solution approach. Section VI presents the experimental results, and Section VII concludes the study.

II. RELATED WORK

Recent studies on quantum optimization can be grouped around three closely related directions: variational hybrid quantum-classical algorithms, QUBO-based reformulation methods, and the encoding of constrained optimization problems for quantum-oriented solvers. McClean et al. introduced the theoretical foundations of variational hybrid quantum-classical algorithms, emphasizing the role of a classical optimizer in training parameterized quantum circuits [1]. Cerezo et al. later provided a comprehensive review of variational quantum algorithms and discussed central challenges, including ansatz design, optimization difficulty, barren plateaus, and measurement cost [2]. These studies show that near-term quantum optimization should be understood as a hybrid computational process rather than a purely quantum procedure. Quantum annealing and adiabatic quantum computation provide another important foundation for optimization. Albash and Lidar reviewed the theoretical background of adiabatic quantum computation and its relationship with quantum annealing [3]. Hauke et al. discussed the practical implementation challenges of quantum annealing, including embedding, hardware limitations, and problem formulation issues[4]. These works are relevant because many quantum annealing systems naturally operate on QUBO or Ising-type formulations. Therefore, transforming an original optimization problem into an appropriate binary quadratic form is a decisive modeling step. QUBO has become a central representation because it connects classical optimization models with quantum-oriented solution methods. Glover et al. presented a detailed tutorial on formulating and using QUBO models, showing how constrained optimization problems can be transformed into unconstrained binary quadratic forms through penalty terms [5]. In the context of gate-based quantum optimization, Hadfield et al. extended the standard Quantum Approximate Optimization Algorithm toward the Quantum Alternating Operator Ansatz for constrained optimization problems [6]. Zhou et al. analyzed the performance and implementation of QAOA on near-term devices [7], while Barkoutsos et al. proposed a CVaR-based strategy to improve variational quantum optimization [8]. These studies indicate that the quality of the problem-to-QUBO mapping strongly affects the performance of both classical and quantum-oriented solvers. Several studies have also focused on reformulation and decomposition strategies. Gambella and Simonetto investigated mixed-binary optimization on classical and quantum computers using multiblock ADMM heuristics [9]. Quintero et al. analyzed QUBO reformulations and penalty characterizations, showing that penalty selection is not merely a numerical detail but a core modeling

decision [10]. More recent studies on inequality constraint encoding further demonstrate that penalty design can significantly influence feasibility and solution quality in QUBO-based optimization [11]. Despite these advances, a complete and reproducible workflow that starts from a nonlinear constrained optimization model, applies variable scaling and binary encoding, embeds constraints via penalty terms, constructs a QUBO model, and compares the decoded solution with that of a classical constrained optimizer remains limited. This study addresses this gap by proposing a transparent QUBO-based reformulation framework and validating it on a nonlinear, resource-constrained production planning problem [12].

III. PROBLEM FORMULATION

Nonlinear constrained optimization problems are generally defined by an objective function to be optimized subject to a set of equality or inequality constraints. In this study, a nonlinear resource-constrained production planning problem is considered as a representative case study. The selected problem is intentionally designed to be simple enough for a QUBO-based reformulation, yet sufficiently expressive to capture nonlinear cost behavior, resource limitations, and bounded decision variables. Let 1 $and\ x\ alt\ 2\ denote$enote the production quantities of two different products. The company aims to maximize its total profit. Each product generates linear revenue; however, production costs increase nonlinearly as production quantity grows. This reflects practical situations in which overtime labor, machine wear, storage limitations, coordination costs, or resource inefficiencies increase as production volume increases.

The profit maximization problem is formulated as follows:

$$\max_{x_1,x_2} P(x_1,x_2) \tag{1}$$

where the nonlinear profit function is defined as:

$$P(x_1,x_2) = r_1x_1 + r_2x_2 - c_1x_1^2 - c_2x_2^2 - c_{12}x_1x_2 \tag{2}$$

In this formulation, $r_1$and $r_2$represent the unit revenues of the products, $c_1$and $c_2$represent nonlinear production cost coefficients, an12 representspresents the interaction cost between the two products. The interaction term may represent shared machinery, production interference, storage competition, or operational complexity caused by producing both products simultaneously.

The problem is subject to resource and capacity constraints:

$$a_{11}x_1 + a_{12}x_2 \le R_1 \tag{3}$$
$$a_{21}x_1 + a_{22}x_2 \le R_2 \tag{4}$$
$$x_1 + x_2 \le R_3 \tag{5}$$
$$0 \le x_1 \le U_1, 0 \le x_2 \le U_2 \tag{6}$$

1 and büyük harf R alt 2 denote, $R_1$and R_2denote the available amounts of two limited resources, such as labor time and raw material. R_3denotes the total production capacity. $U_1$and U_2a2 areupper bounds for the production quantities.

Since most classical optimization algorithms are commonly formulated as minimization procedures, the maximization problem in (1) can be equivalently rewritten as:

$$\min_{x_1,x_2} f(x_1,x_2) = -P(x_1,x_2) \tag{7}$$

Thus, the classical nonlinear constrained optimization model becomes:

$$\min_{x_1,x_2} -r_1x_1 - r_2x_2 + c_1x_1^2 + c_2x_2^2 + c_{12}x_1x_2 \tag{8}$$

subject to the constraints given in (3)–(6).

For the numerical case study, the following parameter values are used:

$$c_1 = 0.08, c_2 = 0.10, c_{12} = 0.03$$
$$2x_1 + 3x_2 \leq 120$$
$$4x_1 + 2x_2 \leq 160$$
$$x_1 + x_2 \leq 70$$
$$0 \leq x_1 \leq 50, 0 \leq x_2 \leq 50$$

Accordingly, the case study problem can be written as:

$$\min_{x_1,x_2} -40x_1 - 55x_2 + 0.08x_1^2 + 0.10x_2^2 + 0.03x_1x_2$$

subject to:

$$2x_1 + 3x_2 \leq 120$$
$$4x_1 + 2x_2 \leq 160$$
$$x_1 + x_2 \leq 70$$
$$0 \leq x_1 \leq 50, 0 \leq x_2 \leq 50$$

This formulation provides a compact nonlinear constrained optimization problem that is suitable for both classical and QUBO-based solution approaches. In the classical setting, the problem can be solved using a constrained optimizer such as Sequential Least Squares Programming. In the QUBO-based setting, the continuous decision variables are first discretized and encoded into binary variables. Then, the inequality constraints are embedded into the objective function using penalty terms. This transformation is described in the next section.

IV. QUBO REFORMULATION FRAMEWORK

Quadratic Unconstrained Binary Optimization is a standard mathematical representation used in many quantum-inspired and quantum-based optimization approaches. A QUBO problem is generally written as:

$$\min_{\mathbf{z}\in\{0,1\}^n} \mathbf{z}^T Q \mathbf{z} \tag{9}$$

where **z**is ais a binary decision vector and büyük harf Qis an n × n real-valued upper triangular or symmetric matrix. The main objective of QUBO reformulation is to transform the original constrained optimization

problem into an unconstrained binary quadratic form. In this study, the transformation is performed through three main steps: variable discretization, binary encoding, and penalty-based constraint embedding.

1. Discretization of Continuous Variables

The original problem includes continuous production variables 1 $and\ x\ alt$ 2.nd $.x_2$. Since QUBO models require binary decision variables, the continuous variables must first be discretized. In this study, production quantities are represented with a unit discretization step. Therefore, each production variable is approximated by a binary expansion:

$$x_1 = \sum_{k=0}^{K_1-1} 2^k\, b_{1k} \tag{10}$$

$$x_2 = \sum_{k=0}^{K_2-1} 2^k\, b_{2k} \tag{11}$$

where $b_{1k}, b_{2k} \in \{0,1\}$are binary variables and K_1, K_2denote the number of binary bits used for representing $x_1$and x_2, respectively. In the numerical case study, six bits are used for each decision variable:

$$K_1 = K_2 = 6 \tag{12}$$

Thus, each variable can represent integer values in the range $[0, 63]$. Although the original upper bounds are $x_1 \leq 50$and $x_2 \leq 50$, the resource constraints further restrict the feasible region. Any infeasible binary solution is discouraged through penalty terms in the QUBO objective function.

2. Slack Variable Representation of Inequality Constraints

The original model contains three inequality constraints. To transform them into QUBO-compatible expressions, each inequality is converted into an equality by introducing a nonnegative slack variable:

$$2x_1 + 3x_2 + s_1 = 120$$
$$4x_1 + 2x_2 + s_2 = 160$$
$$x_1 + x_2 + s_3 = 70$$

where s_1, s_2, and $s_3$are nonnegative slack variables. These variables measure the unused amount of each resource. Since QUBO requires binary variables, the slack variables are also encoded using binary expansions:

$$s_1 = \sum_{k=0}^{K_{s1}-1} 2^k\, y_{1k} \tag{13}$$

$$s_2 = \sum_{k=0}^{K_{s2}-1} 2^k\, y_{2k} \tag{14}$$

$$s_3 = \sum_{k=0}^{K_{s3}-1} 2^k\, y_{3k} \tag{15}$$

where $y_{jk} \in \{0,1\}$are binary slack variables. In this study, the number of bits for the slack variables is selected according to the right-hand side of each constraint. Therefore, $s_1$and $s_3$are represented with seven bits, while $s_2$is represented with eight bits:

$$K_{s1} = 7, K_{s2} = 8, K_{s3} = 7$$

This representation is sufficient to encode the maximum possible slack values for the three resource constraints.

3. Penalty-Based Constraint Embedding

After introducing slack variables, each constraint can be embedded into the objective function using a quadratic penalty term. Let the original minimization objective be:

$$f(x_1, x_2) = -40x_1 - 55x_2 + 0.08x_1^2 + 0.10x_2^2 + 0.03x_1x_2$$

The penalty terms corresponding to the three constraints are defined as:

$$\lambda_1(2x_1 + 3x_2 + s_1 - 120)^2$$
$$\lambda_2(4x_1 + 2x_2 + s_2 - 160)^2$$
$$\lambda_3(x_1 + x_2 + s_3 - 70)^2$$

where λ_1, λ_2, and $\lambda_3$are positive penalty coefficients. These coefficients determine the strength of constraint enforcement. If the penalty coefficients are too small, the optimizer may prefer infeasible solutions with better objective values. If they are too large, the QUBO coefficients may become numerically unbalanced. Therefore, penalty selection is an important modeling decision in QUBO reformulation.

The complete penalized objective function is written as:

$$F(x_1, x_2, s_1, s_2, s_3) = f(x_1, x_2) + \lambda_1(2x_1 + 3x_2 + s_1 - 120)^2 + \lambda_2(4x_1 + 2x_2 + s_2 - 160)^2 + \lambda_3(x_1 + x_2 + s_3 - 7$$

The constrained nonlinear optimization problem is therefore transformed into the following unconstrained binary optimization problem:

$$\min_{\mathbf{z}\in\{0,1\}^n} F(\mathbf{z}) \qquad (16)$$

where all continuous and slack variables are represented by binary variables.

4. Binary Decision Vector

The complete binary vector is constructed by concatenating the binary variables of the original decision variables and the slack variables:

$$\mathbf{z} = [b_{10}, \dots, b_{1,K_1-1}, b_{20}, \dots, b_{2,K_2-1}, y_{10}, \dots, y_{1,K_{s1}-1}, y_{20}, \dots, y_{2,K_{s2}-1}, y_{30}, \dots, y_{3,K_{s3}-1}]^T \qquad (17)$$

For the selected bit configuration, the total number of binary variables is:

$$n = K_1 + K_2 + K_{s1} + K_{s2} + K_{s3}$$
$$n = 6 + 6 + 7 + 8 + 7 = 34$$

Thus, the original two-variable nonlinear constrained optimization problem is transformed into a 34-variable QUBO problem.

5. Construction of the QUBO Matrix

After substituting the binary expansions into the penalized objective function, the resulting expression contains only binary linear and quadratic terms. Since each binary variable satisfies:

$$z_i^2 = z_i \tag{18}$$

all squared binary terms can be represented as linear terms on the diagonal of the QUBO matrix. The final objective can be written in matrix form as:

$$F(\mathbf{z}) = \mathbf{z}^T Q \mathbf{z} + C \tag{19}$$

where $Q$is the QUBO coefficient matrix and $C$is a constant term. Since the constant term does not affect the optimizer, it can be ignored during the solution process. The QUBO problem is therefore expressed as defined in (9)

The QUBO matrix contains the coefficients generated by the original nonlinear objective function and the penalty terms. Diagonal elements represent linear binary contributions, while off-diagonal elements represent pairwise interactions between binary variables.

This reformulation transforms the original nonlinear constrained problem into a QUBO structure that is compatible with quantum annealing, QAOA-based approaches, and classical QUBO solvers. The next section explains the classical and quantum-oriented solution strategies used for solving the constructed QUBO model.

V. QUANTUM SOLUTION APPROACH

After the nonlinear constrained optimization problem is transformed into QUBO form, it becomes suitable for quantum-oriented solution methods. In this study, the constructed QUBO model is considered from three complementary perspectives: classical QUBO search, QUBO-to-Ising representation, and QAOA-based variational quantum optimization. This structure enables comparison of the proposed QUBO reformulation with both classical and quantum-oriented computational paradigms. The QUBO objective was defined in (9), where the binary decision vector and the QUBO coefficient matrix were introduced. In this section, the constructed QUBO model is solved and evaluated using classical and quantum-oriented solution strategies. The solution vector $\mathbf{z}^*$obtained from the QUBO solver is decoded back into the original production quantities $x_1$and x_2, together with the slack variables s_1, s_2, and . s_3. The decoded solution is then evaluated in terms of objective value and constraint feasibility.

1. QUBO-to-Ising Representation

Quantum annealing is one of the most natural computational models for solving QUBO and Ising-type optimization problems. In a quantum annealing-oriented formulation, the QUBO problem is mapped into an energy minimization problem. The goal is to identify a binary configuration corresponding to the system's minimum-energy state.

The QUBO form can be converted into an Ising representation by mapping binary variables z_i ∈ {0, 1} into spin variables s_i ∈ {−1, 1}.$z_i \in \{0,1\}$into spin variables $s_i \in \{-1, +1\}$:

$$z_i = \frac{1 - s_i}{2} \qquad (47)$$

Using this transformation, the QUBO objective can be represented as an Ising Hamiltonian of the following form:

$$H_C = \sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j \qquad (48)$$

where $h_i$represents local fields and J_{ij}represents pairwise interactions between spin variables. In the quantum annealing process, the system evolves from an initial Hamiltonian toward the problem Hamiltonian, and low-energy states are sampled as candidate solutions.

In this study, the quantum annealing-oriented solution is treated as a QUBO-compatible sampling process. The obtained samples are decoded into the original variables, and each solution is evaluated based on the objective value, constraint violations, and feasibility. This enables the QUBO model to be examined under a quantum optimization perspective without requiring an unsupported claim of quantum advantage.

2. Solution Evaluation Criteria

All solution methods are evaluated after decoding the binary vector into the original production variables and slack variables. The comparison is based on objective value, profit value, constraint violation, feasibility, and runtime. This is necessary because QUBO energy alone does not fully indicate the practical quality of a solution. Therefore, each candidate solution is rechecked in the original problem space before being reported in Section VI.

VI. EXPERIMENTAL DESIGN AND RESULTS

The experimental evaluation was designed to test whether the proposed QUBO reformulation preserves the optimal structure of the nonlinear constrained optimization problem defined in (15)–(19). The experiments were implemented in Python using a reproducible workflow. The classical SLSQP optimizer was first used to solve the original problem. Then, the scaled QUBO model was solved using exact enumeration, random search, simulated annealing, and a PennyLane-based QAOA demonstration. For computational validation, the production variables were represented in a scaled integer form as x1 = 10u1 and x2 = 10u2. This reduced representation produces an 18-variable QUBO model, including binary encodings for u1, u2, and the slack variables. This size is sufficient for exact enumeration and also practical for a small-scale PennyLane QAOA demonstration. The QUBO construction follows the binary encoding and penalty-based reformulation framework described in Section IV.

1. Experimental Configuration

The experimental setting is summarized in Table 1. The same objective function and constraints were used across all solution methods. The penalty coefficients were selected as λ1 = λ2 = λ3 = 1000. Each decoded QUBO solution was evaluated again in the original problem space in terms of objective value, profit, feasibility, constraint violation, and runtime.

Table1. Experimental Configuration

Item	**Setting**
Original decision variables	x_1, x_2
Scaled variables	$x_1 = 10u_1, x_2 = 10u_2$
Binary variables for $u_1$and u_2	3 bits each
Slack variables	s_1, s_2, s_3
Total binary variables	18
Penalty coefficients	$\lambda_1 = \lambda_2 = \lambda_3 = 1000$
Classical baseline	SLSQP
QUBO validation	Exact enumeration
Classical QUBO heuristics	Random search, simulated annealing
Quantum-oriented method	PennyLane QAOA demonstration

2. Experimental Results

The obtained results are presented in Table 2. The classical SLSQP optimizer reached x1 = 30 and x2 = 20, corresponding to a maximum profit of 2170. The exact QUBO solver found the same decoded solution. This agreement is the key experimental result of the study, because it confirms that the proposed QUBO reformulation preserves the optimum structure of the original nonlinear constrained optimization problem. The Python/Streamlit implementation used to generate the experimental results is available at https://github.com/cyasar/IHCONCS.

Table 2. Comparison Of Classical, Qubo-Based, And Quantum-Oriented Solution Methods

Method	x_1	x_2	s_1	s_2	s_3	**Objective** $f(x)$	**Profit** $P(x)$	**Violation**	**Runtime**
SLSQP	30.00	20.00	—	—	—	-2170.00	2170.00	1.22×10^{-10}	0.021 s
Exact QUBO	30.00	20.00	0	0	2	-2170.00	2170.00	0.00	9.694 s
Random Search QUBO	10.00	20.00	4	8	4	-1446.00	1446.00	0.00	0.503 s
Simulated Annealing QUBO	40.00	0.00	4	0	3	-1472.00	1472.00	0.00	1.504 s
PennyLane QAOA	20.00	20.00	2	5	4	-1816.00	1816.00	0.00	747.456 s

The small numerical constraint violation reported by SLSQP is due to floating-point precision and is considered zero in practice. The slack variables in Table 2 are reported in the scaled QUBO representation. Random search and simulated annealing generated feasible but suboptimal solutions, with profit values of 1446 and 1472, respectively. The PennyLane QAOA demonstration produced a feasible solution with a profit of 1816, outperforming these two heuristic QUBO methods under the tested settings, but with a substantially higher runtime. Therefore, the QAOA result should be interpreted as a quantum-oriented workflow demonstration rather than evidence of quantum advantage. Overall, the exact QUBO result validates the reformulation, while the PennyLane implementation demonstrates how the constructed QUBO model can be integrated into a hybrid quantum-classical programming environment.

VII. CONCLUSION AND FUTURE WORK

This study presented a QUBO-based reformulation framework for transforming a nonlinear constrained optimization problem into a binary quadratic model. A resource-constrained production planning problem was used as a case study and reformulated through variable scaling, binary encoding, slack variables, and penalty-based constraint embedding. The results showed that the proposed QUBO reformulation preserved the optimum structure of the original problem. The classical SLSQP optimizer and the exact QUBO solver produced the same solution: x1 = 30 and x2 = 20, corresponding to a maximum profit of 2170. This confirms that the transformation was correctly constructed. Other QUBO-based methods produced feasible but suboptimal solutions, while the PennyLane QAOA demonstration showed how the model can be connected to a quantum-oriented workflow.

This study does not claim a quantum advantage. Its main contribution is a transparent and reproducible pathway from nonlinear constrained optimization to QUBO-based optimization. Future work will focus on larger problems, alternative encoding strategies, penalty coefficient selection, QAOA settings, and a Streamlit-based web implementation made available through GitHub at https://github.com/cyasar/IHCONCS.

REFERENCES

[1] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, ‘The theory of variational hybrid quantum-classical algorithms’, *New J. Phys.*, vol. 18, no. 2, p. 023023, Feb. 2016, doi: 10.1088/1367-2630/18/2/023023.

[2] M. Cerezo *et al.*, ‘Variational quantum algorithms’, *Nat. Rev. Phys.*, vol. 3, pp. 625–644, 2021, doi: 10.1038/s42254-021-00348-9.

[3] T. Albash and D. A. Lidar, ‘Adiabatic quantum computation’, *Rev. Mod. Phys.*, vol. 90, no. 1, p. 015002, 2018, doi: 10.1103/RevModPhys. 90.015002.

[4] P. Hauke, H. G. Katzgraber, W. Lechner, H. Nishimori, and W. D. Oliver, ‘Perspectives of quantum annealing: Methods and implementations’, *Rep. Prog. Phys.*, vol. 83, no. 5, p. 054401, 2020, doi: 10.1088/1361-6633/ab85b8.

[5] F. Glover, G. Kochenberger, R. Hennig, and Y. Du, ‘Quantum bridge analytics I: a tutorial on formulating and using QUBO models’, *Ann. Oper. Res.*, vol. 314, no. 1, pp. 141–183, Jul. 2022, doi: 10.1007/s10479-022-04634-2.

[6] S. Hadfield, Z. Wang, B. O’Gorman, E. G. Rieffel, D. Venturelli, and R. Biswas, ‘From the Quantum Approximate Optimization Algorithm to a Quantum Alternating Operator Ansatz’, *Algorithms*, vol. 12, no. 2, p. 34, Feb. 2019, doi: 10.3390/a12020034.

[7] L. Zhou, S.-T. Wang, S. Choi, H. Pichler, and M. D. Lukin, ‘Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices’, *Phys. Rev. X*, vol. 10, no. 2, p. 021067, 2020, doi: 10.1103/PhysRevX.10.021067.



[8] P. Kl. Barkoutsos, G. Nannicini, A. Robert, I. Tavernelli, and S. Woerner, 'Improving variational quantum optimization using CVaR', *Quantum*, vol. 4, p. 256, 2020, doi: 10.22331/q-2020-04-20-256.
[9] C. Gambella and A. Simonetto, 'Multi-block ADMM heuristics for mixed-binary optimization on classical and quantum computers'. 2020. doi: 10.48550/arXiv.2001.02069.
[10] R. Quintero, D. Bernal, T. Terlaky, and L. F. Zuluaga, 'Characterization of QUBO reformulations for the maximum k-colorable subgraph problem'. 2021. doi: 10.48550/arXiv.2101.09462.
[11] A. Montanez-Barrera, D. Willsch, A. Maldonado-Romo, and K. Michielsen, 'Unbalanced penalization: A new approach to encode inequality constraints of combinatorial problems for quantum optimization algorithms', *Quantum Sci. Technol.*, vol. 9, no. 2, 2024, doi: 10.1088/2058-9565/ad35e4.
[12] F. W. Glover and G. A. Kochenberger, 'A Tutorial on Formulating QUBO Models', *CoRR*, vol. abs/1811.11538, 2018, [Online]. Available: http://arxiv.org/abs/1811.11538

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Long Distance Environmental and Security System

Mohammad Mohatram
Electrical and Electronics Engineering
Global College of Engineering and Technology
Muscat, Oman
m.mohatram@gcet.edu.om
ORCID:0000-0002-5459-5631

Muhammad Irfan Memon
Electrical and Electronics Engineering
Global College of Engineering and Technology
Muscat, Oman
m.memon@gcet.edu.om
ORCID:0000-0003-4194-3657

Mohammed Khalfan Mohammed Al Hudaifi
Electrical and Electronics Engineering
Global College of Engineering and Technology
Muscat, Oman
20241107@gcet.edu.om

Javeed Hussain Shaik
Electrical and Electronics Engineering
Global College of Engineering and Technology
Muscat, Oman
s.javeedhussain@gcet.edu.om
ORCID:0000-0002-2696-1163

Abstract
This paper proposed a hybrid long-distance monitoring system that integrates a Wireless Sensor Network (WSN) and Internet of Things (IoT) to overcome the connectivity issues in remote locations. The proposed system uses an Arduino Nano microcontroller for data collection on the remote end and an HC-12 radio module to enable long-range wireless communication without an internet connection. The sensor data, such as temperature and humidity from a DHT11 sensor and fire detection from a flame sensor, are gathered and transmitted wirelessly over a range of 100 meters in an open environment. The NodeMCU (ESP8266) module on the receiving side receives incoming data, displays it on a 16x2 I2C LCD screen, and transmits it to the Blynk cloud platform, where it is displayed in real time. Experimental evaluation demonstrated reliable wireless communication over 100 m with a 95% packet delivery ratio, RF latency below 2 s, and mobile alerts delivered within 5s.

Keywords: wireless sensor networks, IOT, HC-12, environmental monitoring, ESP8266, remote systems, smart security.

I. INTRODUCTION

The Internet of Things (IoT) has transformed how humans interact with their environment, enabling data collection, analysis, and relay with minimal human intervention [1]. IoT-based monitoring systems are common in agriculture, industry, smart homes, and healthcare. Embedded systems, wireless communication, and cloud computing have grown rapidly, making monitoring systems cost-effective. However, these technologies are largely ineffective in remote or rural areas such as farmland, forests, mountains, and coastal borders [2]. Wi-Fi has a limited range (~100 m), while GSM-based systems require recurring subscription fees and network infrastructure. Neither operates reliably when infrastructure is damaged or unavailable. A low-cost monitoring system that works over long distances without internet dependency is increasingly needed, especially in developing countries where digital infrastructure is lacking but environmental and security monitoring is critical. Although numerous IoT monitoring systems have been proposed, most rely on continuous Internet connectivity or gateway infrastructure. Consequently, these systems cannot operate effectively in remote regions where communication infrastructure is unavailable. Moreover, few studies integrate long-range RF communication with cloud-based monitoring while simultaneously supporting environmental sensing and security monitoring.

This research addresses these limitations through a hybrid HC-12–IoT architecture. The principal contributions of this work are:

i) create an effective long-range wireless sensing system;
ii) combine multiple environmental and security sensors;
iii) provide real-time alerts and data visualization via LCD and mobile app; and
iv) reduce reliance on constant internet connectivity for remote deployment [3].

II. LITERATURE REVIEW

Recent years have seen significant research in IoT-based environmental monitoring. Cloud-connected sensor systems monitoring soil moisture, temperature, and humidity have enhanced agricultural productivity [4]. Smart greenhouse systems using Arduino and cloud computing have reduced manual monitoring by over sixty percent [5]. ESP32-based greenhouse systems enable real-time data collection and automated control. IoT-based air pollution monitoring systems combine environmental sensors with cloud visualization. ESP32-based cloud-enabled IoT systems using Firebase show high transmission reliability and low latency. Wireless Sensor Networks (WSNs) consist of spatially distributed sensor nodes that gather and wirelessly transmit data to a central location. Recent studies confirm the importance of smart energy management in WSNs. Kumar et al. proposed clustering and routing algorithms that enhance network throughput and lifetime [6]. Various wireless communication technologies have been evaluated for remote IoT monitoring. Wi-Fi is range-limited. GSM/GPRS requires subscription fees and may be inaccessible in remote areas. Zigbee is low-power but short-range. LoRa and LoRaWAN provide long-range communication with low power consumption. They are effective in rural deployments with a multi-kilometer range. However, LoRaWAN requires gateways and infrastructure, increasing cost and complexity [7], [8]. Recent studies have increasingly adopted hybrid communication architectures to improve the reliability and scalability of IoT monitoring systems [9]. Abdulmalek et al. [10] implemented a hybrid Wi-Fi/LoRaWAN architecture for remote healthcare monitoring, demonstrating reliable long-range communication while relying on gateway infrastructure. Similarly, Hassan et al. [11] proposed a hybrid fog–cloud framework for structural health monitoring that improves processing efficiency but depends on continuous network connectivity. In contrast, the proposed system employs HC-12 serial RF communication to provide infrastructure-independent local sensing while maintaining cloud accessibility through the receiver node. The HC-12 serial radio module (433 MHz band) is a low-cost, straightforward wireless communication device offering up to 1 kilometer range in open space without gateways or internet access. It is simpler to implement than LoRa and more cost-effective for small-to-medium-range deployments that prioritize infrastructure independence [12]. Despite rich literature on IoT monitoring, most deployed systems remain overly reliant on uninterrupted internet connectivity. Offline IoT operation frameworks show that connectivity losses significantly degrade functionality, a major weakness for remote deployments [13]. Hybrid systems combining offline wireless communication with cloud-based monitoring are rare [14]. Most solutions are domain-focused and do not combine environmental and security sensors into a unified platform [15]. Table 1 presents a comparison of wireless communication technologies.

Technology	**Range**	**Infrastructure Required**	**Cost**
Wi-Fi	~100 m	Yes	Low
Zigbee	~100–300 m	Optional	Medium

GSM/NB-IoT	Several km	Cellular network	High
LoRa	Several km	Gateway required	Medium–High
HC-12	~100 m tested	No	Low

Table 1. Wireless Communication Technologies Comparison

This work addresses these limitations by integrating the HC-12 module with IoT cloud platforms to create a cost-effective, multi-sensor hybrid monitoring system that operates in both offline and online modes. Figure 1 compares the wireless communication technologies considered in this study.

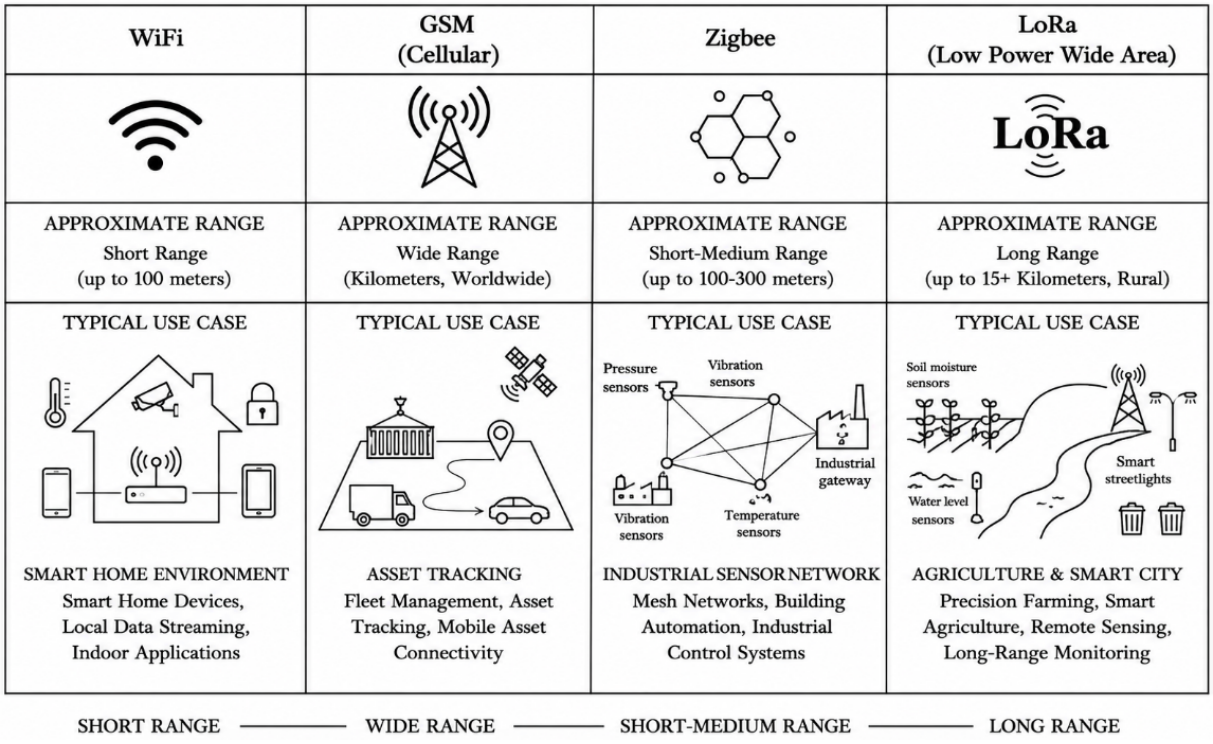


Figure 1. Comparison of communication technologies in IoT

III. METHODOLOGY

1. System Architecture

A dual-node architecture was implemented, comprising a Transmitter Unit and a Receiver Unit that communicate via HC-12 RF serial modules operating at 433 MHz. The transmitter acquires environmental data and forwards it wirelessly; the receiver processes, displays, and uploads the data to the cloud. The complete circuit implementation is shown in Fig. 2.

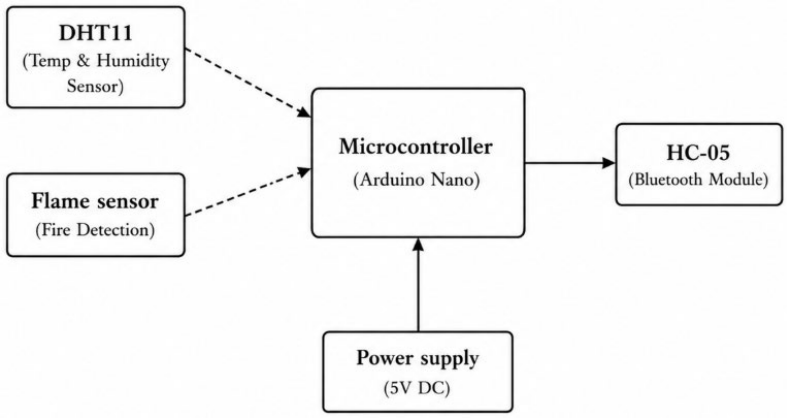


Figure 2. Block diagram of the proposed system

2. System Hardware Implementation

The proposed system is implemented using Arduino Nano and NodeMCU (ESP8266) microcontrollers. The Arduino Nano acquires temperature, humidity, and flame data from the connected sensors, formats the measurements into serial packets, and transmits them through an HC-12 wireless module operating at 433 MHz. The NodeMCU receives the transmitted data, displays real-time measurements on a 16 × 2 I²C LCD, communicates with the Blynk IoT platform for cloud monitoring and notifications, and manages system alerts. Table 2 summarizes the hardware components and their functions.

Table 2. Hardware Components of the Proposed System

Component	**Function**
Arduino Nano	Collects sensor data, processes measurements, formats data packets, and controls wireless transmission.
NodeMCU (ESP8266)	Processes received data, updates the LCD, communicates with the Blynk IoT platform, and generates alerts.
DHT11 Sensor	Measures ambient temperature and relative humidity.
Flame Sensor	Detects the presence of flame and provides a digital detection signal.
HC-12 Wireless Module	Enables 433 MHz RF communication between transmitter and receiver.
16 × 2 I²C LCD	Displays real-time temperature, humidity, flame status, and system messages.
Blynk IoT Platform	Provides cloud monitoring, data logging, and mobile notifications.
Li-ion Battery + TP4056 + DC–DC Boost Converter	Provides a regulated 5 V power supply for system operation.

The Arduino Nano was selected for its low cost, sufficient GPIO pins, and ease of programming. The ESP8266 was selected because of its integrated Wi-Fi capability for direct cloud communication. HC-12 was preferred over LoRa due to its lower implementation complexity and the absence of gateway requirements in short- to medium-range deployments. The complete hardware configuration is summarized in Table 2. Both units are powered by lithium-ion batteries with TP4056 charging modules and DC-DC boost converters to provide regulated 5 V operation.

3. Circuit Design

The complete circuit was modeled and simulated in Proteus 8 Professional before hardware implementation, enabling verification of component interconnections and signal behavior. The finalized circuit diagram is shown in Fig. 3.

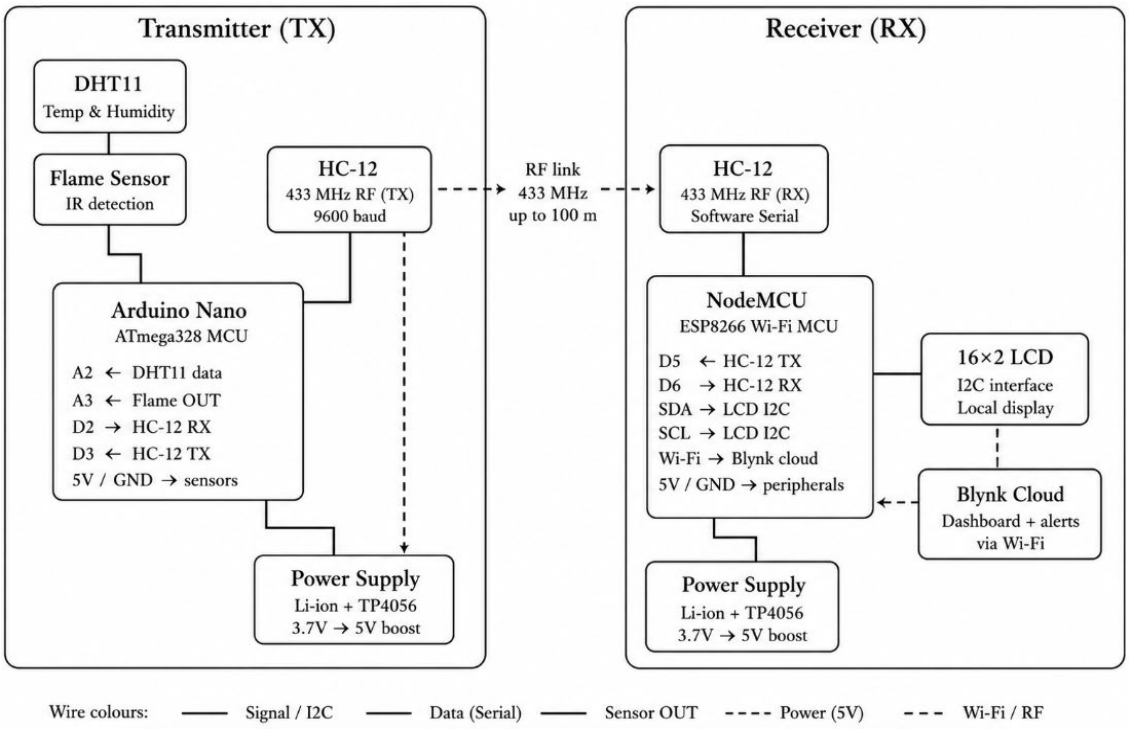


Figure 3. Circuit diagram of the proposed system

Key wiring configurations are as follows. The DHT11 sensor is connected to Arduino Nano pin A2 with a pull-up resistor. The flame sensor digital output is connected to pin A3 for HIGH/LOW detection. The transmitter-side HC-12 TX and RX lines are connected to Arduino pins 2 and 3, respectively, while the receiver-side HC-12 is connected to NodeMCU pins D5 and D6. The LCD is interfaced via the I2C protocol using the SDA and SCL lines, minimizing wiring complexity.

4. Firmware Development

Continuous polling loop (1-second frequency) reading sensors and Firmware for both nodes was developed in C++ using the Arduino IDE. The transmitter executes a continuous polling loop at one-second intervals, reading the DHT11 and flame sensor outputs and transmitting comma-delimited data strings via the HC-12 module. The receiver parses incoming data packets, updates the LCD, and forwards readings to the Blynk platform at 30-second intervals to avoid exceeding API rate limits. Automated alerts are triggered under two conditions: flame detection, indicated by a LOW signal from the flame sensor, and temperature readings exceeding 30 °C. In both cases, a push notification is dispatched to the operator's registered mobile device using the Blynk.notify() function. The firmware logic for both nodes is illustrated in Fig. 4.

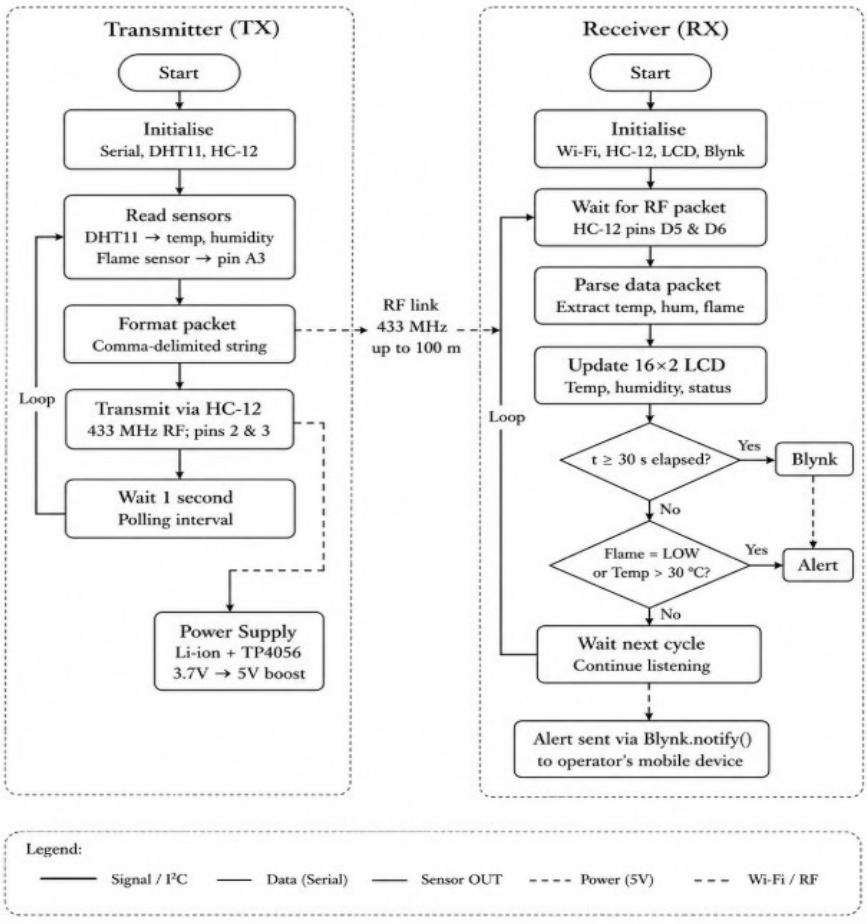


Figure 4. Filmware flowchart -transmitter node and receiver node

IV. RESULTS AND DISCUSSION

1. Hardware Implementation

Both transmitter and receiver units were successfully assembled and tested. The DC-DC boost converter maintained a regulated output of 4.9-5.1 V under load, confirming stable power delivery to all components. A

pin assignment correction (flame sensor from A1 to A3) was applied without operational disruption, demonstrating the flexibility of the Arduino Nano platform. The I2C LCD interface operated reliably throughout testing, with no communication faults observed. Figure 5 presents the prototype of the implemented system.



Figure 5. Prototype of the implemented system (transmitter and receiver units)

Table 3. Sensor Performance Characteristics

Parameter	Observed Range	Remarks
Temperature (DHT11)	0°C - 50°C (±2°C)	Within specifications
Humidity (DHT11)	20% - 90% RH (±5% RH)	Within specifications
Flame detection	Up to ~100 cm	Sensitivity adjustable

2. Wireless Communication Performance

The HC-12 modules operated reliably at 9600 baud throughout all test scenarios. Open-field range exceeded 100 m with stable packet delivery, while indoor range was reduced to 30-50 m due to wall attenuation. These results confirm the suitability of 433 MHz RF communication for short- to medium-range environmental monitoring deployments. The packet delivery ratio decreased slightly with increasing distance due to RF attenuation and environmental interference. Nevertheless, the system maintained over 95% successful transmission at 100 m, demonstrating adequate reliability for remote monitoring applications.

Table 4. HC-12 Communication Performance at Different Distances and Environments

Distance (m)	Environment	Packet Delivery Ratio (%)	Latency
10	Indoor	100	<1 sec
30	Indoor	98	<1 sec
50	Indoor	92	1-2 sec
50	Outdoor	100	<1 sec
100	Outdoor	95	1-2 sec

3. Signal Loss Detection

The 10-second watchdog timeout mechanism functioned correctly under all test conditions. Upon deliberate power-off of the transmitter unit, the receiver correctly detected the absence of incoming packets and displayed a “signal loss” warning on the LCD within 10 seconds. Normal operation resumed automatically upon transmitter recovery, with no manual intervention required. This behavior confirms the robustness of the supervisory logic for unattended deployment scenarios.

4. Alert and Notification System

All automated alert conditions were verified under controlled laboratory conditions. Push notification latency remained below 5 seconds under stable Wi-Fi connectivity, meeting the real-time alerting requirement. Figure 6 presents the developed Blynk dashboard used for real-time monitoring.

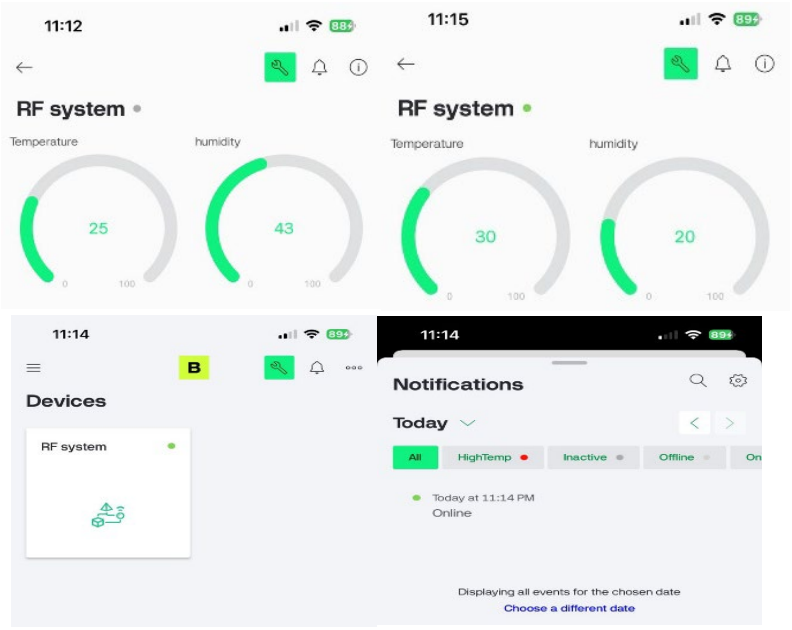


Figure 6. Blynk IoT dashboard showing real-time sensor data and alert notification

- Fire detection: Flame sensor (LOW signal) triggered Blynk push notification “Fire Detected” and LCD warning. Respond immediately.
- High temperature alert: Temperature threshold greater than 30 °C triggered email notification via Blynk. Threshold configurable in firmware.
- Notification latency: Less than 5 seconds under stable Wi-Fi conditions. Delivered via Blynk.notify() to registered mobile device.

5. Cloud Integration

The Blynk cloud platform successfully received temperature, humidity, and flame status data at 30-second upload intervals across all test sessions. Historical data logging functioned correctly, enabling retrospective

analysis of environmental conditions. The 30-second interval was found to strike an appropriate balance between data granularity and API rate-limit compliance, with no upload failures recorded during extended operation.

6. Power System Performance

The TP4056 charging modules with DC-DC boost converters reliably delivered 5 V to both nodes throughout continuous operation testing. The receiver exhibited a shorter runtime due to the higher current consumption caused by the ESP8266's continuous Wi-Fi radio activity. The power profile is compatible with small photovoltaic charging systems, making the design suitable for remote outdoor deployments where mains power is unavailable.

7. Overall system performance summary

Table 5 consolidates the key measured performance metrics across all subsystems.

Table 5. Overall System Performance Metrics

Parameter	**Measured Result**
Open-field RF range	~100 m
Indoor RF range	30–50 m
Average RF latency	<2 sec
Blynk notification latency	<5 sec
Packet delivery ratio (100 m outdoor)	95%
Flame detection distance	~100 cm
Cloud upload interval	30 sec
Transmitter battery runtime	~6–8 hours
Receiver battery runtime	~5–7 hours

V. CONCLUSION

This paper presented a low-cost hybrid WSN-IoT environmental and security monitoring system combining infrastructure-free HC-12 RF communication with Blynk cloud connectivity. The system reliably transmits temperature, humidity, and flame-detection data over 100 m without internet connectivity, triggers mobile push notifications within 5 s of a hazard event, and automatically detects signal loss via a hardware-independent watchdog. At a total component cost below USD 35 and battery autonomy of 6-8 h per charge, the platform is well-suited to remote agricultural, border-security, and infrastructure-monitoring deployments in resource-constrained environments. Future work will target: (i) substituting HC-12 with LoRa for extended range; (ii) CRC-based packet validation and retransmission for higher reliability; (iii) integrating gas, smoke, and motion sensors for broader security coverage; (iv) solar-powered continuous operation; (v) a multi-node architecture for simultaneous multi-site monitoring; and (vi) edge-based machine-learning anomaly detection.

REFERENCES

[1]. M. Mansour et al., "Internet of Things: A Comprehensive Overview on Protocols, Architectures, Technologies, Simulation Tools, and Future Directions," Energies, vol. 16, no. 8, p. 3465, 2023.

[2]. H. Singh et al., "An IoT-based smart home automation system using a sensor node," in 2018 4th International Conference on Recent Advances in Information Technology (RAIT), IEEE, 2018.
[3]. N. Gondchawar and R. Kawitkar, "IoT-based smart agriculture," International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 6, pp. 838-842, 2016.
[4]. A. Ben Sada et al., "A context-aware edge computing framework for smart internet of things," Future Internet, vol. 15, no. 5, p. 154, 2023.
[5]. Y. Song, J. Bi, and X. Wang, "Design and implementation of an intelligent monitoring system for the agricultural environment in IoT," Internet of Things, vol. 25, p. 101029, 2024.
[6]. A. Kumar, N. Gaur, and A. Nanthaamornphong, "Wireless optimization for sensor networks using IoT-based clustering and routing algorithms," PeerJ Computer Science, vol. 10, p. e2132, 2024.
[7]. J. Bravo-Arrabal et al., "Realistic deployment of hybrid wireless sensor networks based on ZigBee and LoRa for search and rescue applications," IEEE Access, vol. 10, pp. 64618-64637, 2022.
[8]. Alumfareh, M. F., et al. "An Intelligent LoRaWAN-based IoT Device for Monitoring and Control Solutions in Smart Farming through Anomaly Detection Integrated with Unsupervised Machine Learning." IEEE Access, vol. 12, 2024.
[9]. Andrade, A., et al. "IoT-Based Vital Sign Monitoring: A Literature Review." Smart Health, vol. 32, 2024, Article 100462.
[10]. Abdulmalek, S., et al. "LoRaWAN-Based Hybrid Internet of Wearable Things System Implementation for Smart Healthcare." Internet of Things, vol. 25, 2024, Article 101124.
[11]. Hassan, M. et al. "Optimizing Structural Health Monitoring Systems Through Integrated Fog and Cloud Computing Within IoT Framework." IEEE Access, vol. 12, 2024.
[12]. A. Abubakar, M. T. El-Gammal, and A. A. Alarood, "End-To-End Fully-Informed Network Nodes Associated with 433 MHz Outdoor Propagation Environment," International Journal of Computing and Digital Systems, vol. 10, pp. 1359-1373, 2020.
[13]. C.-Y. Wu and K.-H. Huang, "A framework for offline operation of smart and traditional devices of IoT services," Sensors, vol. 20, no. 21, p. 6012, 2020.
[14]. O. Ali et al., "A comprehensive review of Internet of Things: Technology stack, middleware, and fog/edge computing interface," Sensors, vol. 22, no. 3, p. 995, 2022.
[15]. M. A. Khan and K. Salah, "IoT security: Review, blockchain solutions, and open challenges," Future Generation Computer Systems, vol. 82, pp. 395-411, 2018.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Comparative Evaluation of Heuristic Search Algorithms for Reward-Constrained Museum Visitor Routing on a Real-World Graph

Petar Mileski
Department of CSE
University of Information Science and Technology St. Paul the Apostle
Ohrid, North Macedonia
petar.mileski@cse.uist.edu.mk
ORCID: 0009-0001-7120-9763

Dijana Capeska Bogatinoska
Department of ISVMA
University of Information Science & Technology St. Paul the Apostle
Ohrid, North Macedonia
dijana.c.bogatinoska@uist.edu.mk
ORCID: 0000-0001-5474-8290

Abstract
Optimizing visitor routes in museums is a challenging reward-constrained path-planning problem in which visitors aim to maximize exhibit exposure within a limited time budget. We formulate this as a reward-maximization problem on an undirected graph model of the Museum of Modern Art (MoMA) 5th floor (30 nodes, 24 galleries, 465 total exhibits) and systematically compare five search algorithms: Breadth-First Search (BFS), Depth-First Search (DFS), Greedy Best-First Search, A*, and a preliminary hybrid variant (Hybrid-ADFS). Experiments vary in time budgets (10/30/60/90 minutes), starting locations, reward distributions, and exit constraints. A* achieves complete room coverage (465 exhibits, 100% at budgets 60/90 min) with up to 2.38 million node expansions. Greedy plateaus at 199 exhibits with near-constant expansions (52 nodes). Uninformed methods (BFS/DFS) yield low exposure (72 exhibits) regardless of budget. Under realistic crowd-dependent and time-decaying rewards, the performance gap narrows dramatically: Greedy reaches nearly the same reward ceiling as A* (242 vs. 243 at 90 min) despite orders-of-magnitude fewer expansions. These findings contradict the assumption that more sophisticated algorithms are always better and highlight that simple greedy strategies become nearly as effective as computationally expensive informed methods under practical visitor constraints.

Keywords: heuristic search, museum visitor routing, reward maximization, A* algorithm, greedy search, graph-based planning

I. INTRODUCTION

Museums worldwide face the ongoing challenge of guiding visitors through large collections within limited time budgets while maximizing exposure to high-value exhibits. With the rise of mobile applications offering location-based content, personalized routing has become feasible [1], [2], [3], yet few studies systematically evaluate traditional search algorithms on real museum layouts under realistic constraints such as step budgets and discrete exhibit rewards [4], [5], [6]. This work addresses this gap by treating museum visitor routing as a reward-constrained path-optimization problem on an undirected graph derived from the MoMA 5th-floor permanent collection layout.

1. Research Questions

RQ1: Which heuristic search algorithms achieve optimal or near-optimal solutions for reward-constrained museum visitor routing under varying step budgets?
RQ2: What are the computational trade-offs (execution time, memory usage, node expansions) between solution quality and resource consumption across the evaluated algorithms?

RQ3: How do algorithm performance characteristics vary across different starting locations, reward distributions, and exit constraints?

2. Contributions

(1) A clear mathematical formulation of museum visitor routing as a reward-constrained optimization problem on a realistic graph model. **(2)** Systematic comparative evaluation of five search algorithms (BFS, DFS, Greedy, A*, and a preliminary hybrid variant) under realistic constraints. **(3)** Experimental validation using a validated graph of the MoMA 5th floor (30 nodes, 24 galleries, 465 total exhibits). **(4)** Detailed performance analysis, including statistical comparison (pairwise t-test), providing practical algorithm selection guidelines. **(5)** Demonstration that spatial clustering of high-value exhibits dominates routing decisions, with implications for museum design and visitor-flow optimization.

II. RELATED WORK

The problem of routing visitors in a museum has been modeled as variants of the Orienteering Problem (OP) to optimize routes that maximize "profit" subject to given time or distance constraints [7-10]. Research has included the Team Orienteering Problem with Time Windows (TOPTW) for personalized tourist routing [9], which was later extended to include time-dependent links and mandatory visitation [10]. Heuristic and metaheuristic algorithms prevail due to the NP-hardness of the problem, including genetic algorithms [1], other metaheuristics [11], [12], iterated local search [9], and ant colony optimization applied to multi-day tourist itineraries. Museum-specific projects include probabilistic tour-guide robots such as Minerva, which combine indoor navigation with interactive visitor guidance [13]. Indoor navigation based on A* and greedy algorithms is widely used in robotics and mobile guides [5], [13]. Recent work has explored reinforcement learning for reward-based path planning in dynamic indoor environments [14], [15], and transformer-based models for personalized itinerary recommendation [16], representing the modern direction of this research area. No studies directly compare BFS, DFS, Greedy, or A* on real-life museum graphs with rewards. This paper addresses that gap through a comparative case study on the MoMA 5th floor.

III. MATHEMATICAL PROBLEM FORMULATION

Given a museum floor represented as an undirected graph, a fixed time budget, and a set of galleries each carrying a discrete reward (exhibit count), the problem is: find a walk starting from a designated entrance, ending at a valid exit, within the step budget, that maximizes the total reward collected. Formally, the inputs are the graph $G = (V, E)$, a start node s, a step budget B, a reward function R, and a set of valid exits T. The output is a path P from s to some $t \in T$ such that $\text{length}(P) \leq B$ and the sum of $R(v_i)$ for all visited galleries v_i is maximized. Each gallery's reward is collected at most once, regardless of revisits. We model the museum as an undirected graph $G = (V, E)$, where V is the set of vertices (gallery rooms, hallways, entrance/exit points), E is the set of edges representing direct connections (each edge has unit weight, corresponding to one navigation step), $V_g \subseteq V$ is the subset of gallery vertices, and $R: V_g \rightarrow R \geq 0$ is the reward function assigning each gallery a non-negative value representing its exhibit count (total reward across all galleries = 465). A visitor starts at an entrance $s \in S \subseteq V$ and must reach any exit $t \in T \subseteq V$ in at most B steps. Paths may revisit vertices, but reward is collected only once per gallery. The objective is to maximize the total collected reward: $\max \Sigma R(v_i)$ for $v_i \in P \cap V_g$, subject to $\text{length}(P) \leq B$, where P is a path from s to $t \in T$. This formulation is a variant of the Orienteering Problem (OP), which is NP-hard even on planar graphs [7-9].

IV. ALGORITHMS

We evaluate five search algorithms: BFS and DFS as uninformed baselines, Greedy Best-First Search as a heuristic approach, A* as an optimal informed search, and a preliminary Hybrid A*-DFS (Hybrid-ADFS) variant attempting to balance exploration and efficiency.

1. Breadth-First Search (BFS)

BFS explores levels from the start node, guaranteeing the shortest path to any reachable exit. Time Complexity: O(|V| + |E|). Space Complexity: O(|V|). Expected performance: finds the shortest path quickly but ignores reward, yielding low total exposure.

2. Depth-First Search (DFS)

DFS explores as far as possible along each branch before backtracking. Time Complexity: O(|V| + |E|). Space Complexity: O(|V|). Expected performance: may find higher-reward paths than BFS but remains reward-agnostic.

3. Greedy Best-First Search

Greedy prioritizes nodes with the highest immediate reward using a priority queue. Time Complexity: O(|E| log |V|) with a binary heap. Space Complexity: O(|V|). Expected performance: rapidly converges to a local optimum with low expansions but may miss the global maximum.

4. A* Search

A* uses an admissible heuristic (precomputed distance to the nearest exit) to guide search toward exits while maximizing collected reward along the path [17-19]. Time Complexity: O(b^d) worst case. Space Complexity: O(b^d). Expected performance: finds near-optimal reward paths efficiently, though node expansions grow exponentially.

5. Preliminary Hybrid Variant (Hybrid-ADFS)

Hybrid-ADFS combines A* with periodic depth-limited DFS to reduce the number of expansions at extended budgets. While empirical results show it did not improve reward collection over baselines under the tested parameters, Hybrid-ADFS is included in this study for two reasons. First, it represents a natural extension of combining global informed search with local depth-limited exploration, a direction with theoretical motivation for large graphs where A* becomes memory-prohibitive. Second, its underperformance provides a useful negative result: simply interleaving A* and DFS without reward-aware branching yields no gains, which informs future hybrid designs. Refinement through adaptive parameter tuning and reward-guided DFS branching remains a direction for future work.

6. Experimental Design

Experiments were conducted on a desktop computer with an Intel Core i5 processor and 16 GB of RAM, running Python 3.13. The graph is derived from the publicly available MoMA 5th-floor layout: 30 nodes (24 galleries, 3 hallways, 3 exits), edges representing direct connections, and rewards based on actual exhibit counts

(total: 465). Reward variants tested: original (real counts), uniform (10 per gallery), and skewed (doubled for iconic rooms).

Experimental parameters: step budgets B ∈ {10, 30, 60, 90} minutes (1 step ≈ 1 minute walking time); starting locations: three elevator entrances (H_NE, H_E, H_SW); exit constraints: all exits, only_SW, only_NE, only_E; single deterministic run per configuration. Performance metrics: Total Exhibit Reward (TER), Reward Per Step (RPS), Node Expansions (NE), Execution Time (ET), Peak Memory Usage (MU), Room Coverage Ratio (RCR), Exit Used (ESD), Backtracking Ratio (BR), and Path Length.

V. RESULTS

We evaluated the five algorithms across budgets of 10, 30, 60, and 90 minutes, three starting locations, three reward distributions, and four exit constraints. Table 1 summarizes the paired t-test results for A* vs. Greedy at budget 30 (the representative case where differences are clear but A* has not yet reached full coverage). Statistical tests confirm that differences in reward performance reflect systematic algorithmic behavior rather than random variation.

1. Statistical Significance Testing

Paired t-tests on TER across three starting locations for each budget ($\alpha = 0.05$) confirm: A* significantly outperforms Greedy for all tested budgets ($p < 0.05$), with the largest differences at longer budgets. Greedy shows no significant improvement beyond budget 30 ($p > 0.05$), confirming its plateau effect. Uninformed methods (BFS/DFS) are consistently outperformed by both informed approaches ($p < 0.001$). Hybrid-ADFS showed no improvement over uninformed baselines ($p > 0.05$).

Table 1. Paired t-test A* vs. Greedy at budget 30 (mean TER ± std dev)

Metric	**A***	**Greedy Best-First**
Mean TER	457 ± 3.0	156.7 ± 63.0
95% Confidence Interval	[449.97, 464.03]	[7.95, 305.39]
Paired t-statistic	6.77	–
p-value	0.021	–
Significance	Yes ($p < 0.05$)	–

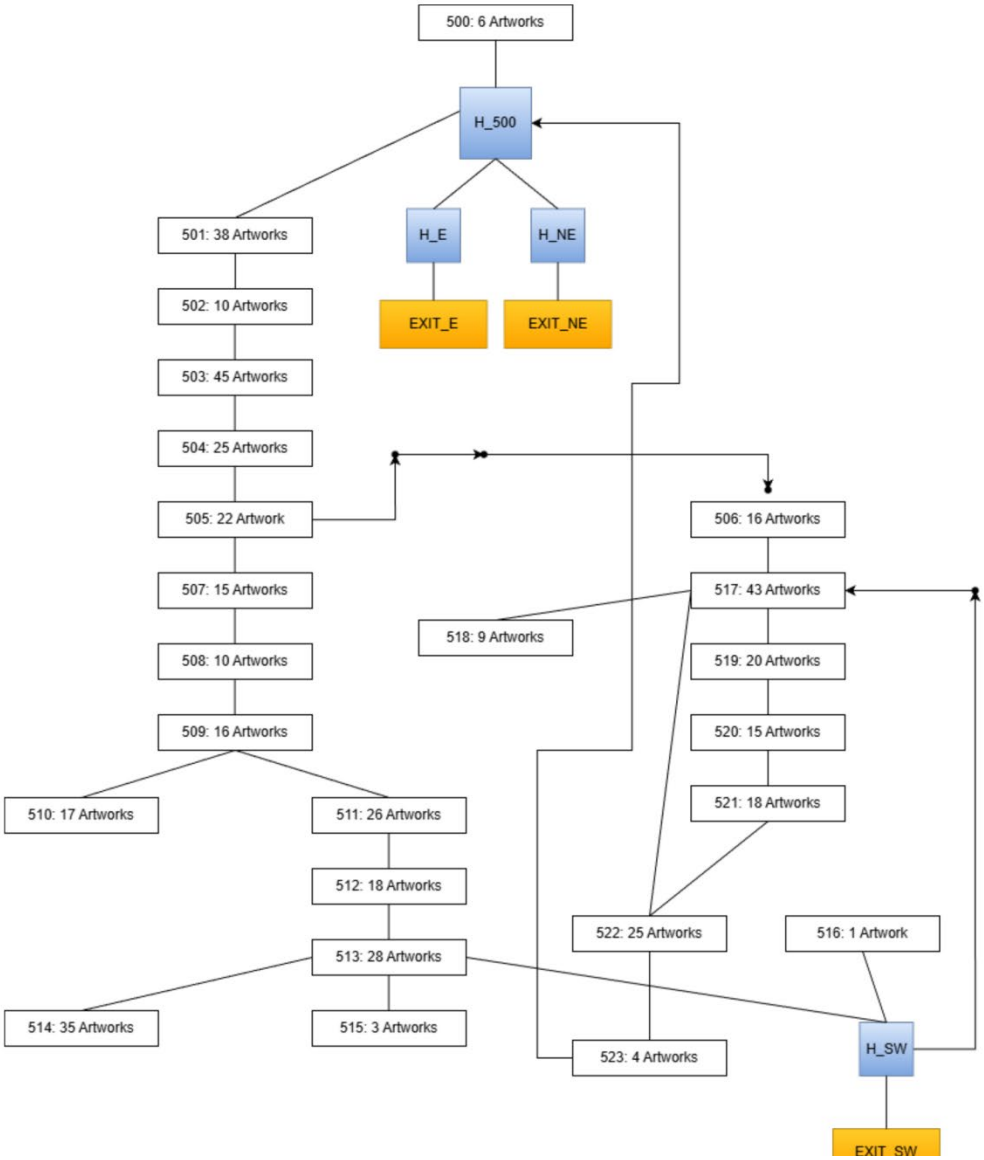


Figure 1. Abstracted graph representation of the MoMA 5th floor.

2. Algorithm Performance Comparison

A* dominates in reward and coverage, but at a significantly higher computational cost. Greedy provides the best balance for practical use. Uninformed and preliminary hybrid methods remained limited in reward collection. Table 2 summarizes the overall average performance across all budgets (10–90 minutes) and starting locations.

Table 2. Algorithm performance comparison (averaged across all budgets and starts)

Algorithm	Avg Reward (mean ± std)	Exec Time (s)	Nodes Explored
BFS	72.0 ± 0.0	0.0003	31
DFS	72.0 ± 0.0	0.0006	41
Greedy	156.7 ± 63.0	0.0009	52
A*	396.5 ± 127.5	58.5	2.38M
Hybrid-ADFS	72.0 ± 0.0	0.036	43

3. Starting Location, Reward, and Exit Impact

Starting location significantly affects uninformed and Greedy performance under short budgets (early termination near exits), but A* remains robust across all starting locations. Reward distributions (original, uniform, skewed) and exit constraints produce negligible differences in TER and path selection, confirming that spatial clustering dominates routing behavior. Figures 2–4 present the key performance metrics and an A* representative path.

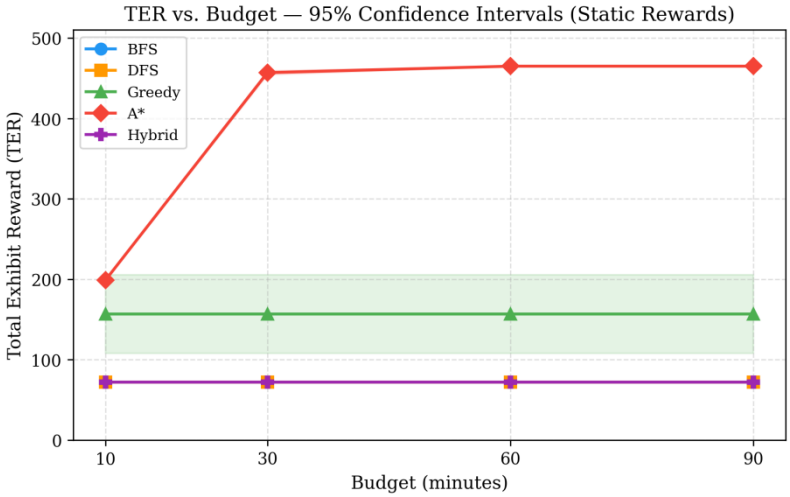


Figure 2. Total Exhibit Reward (TER) vs. step budget with 95% Confidence Intervals

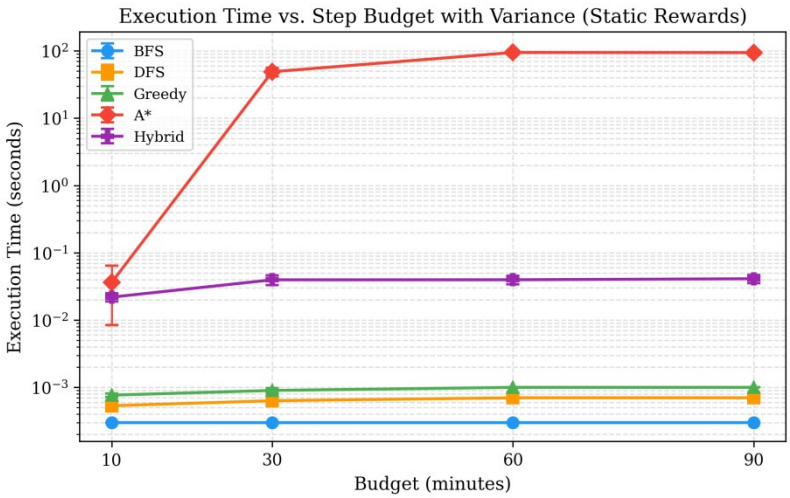


Figure 3. Execution time vs. step budget (log scale)

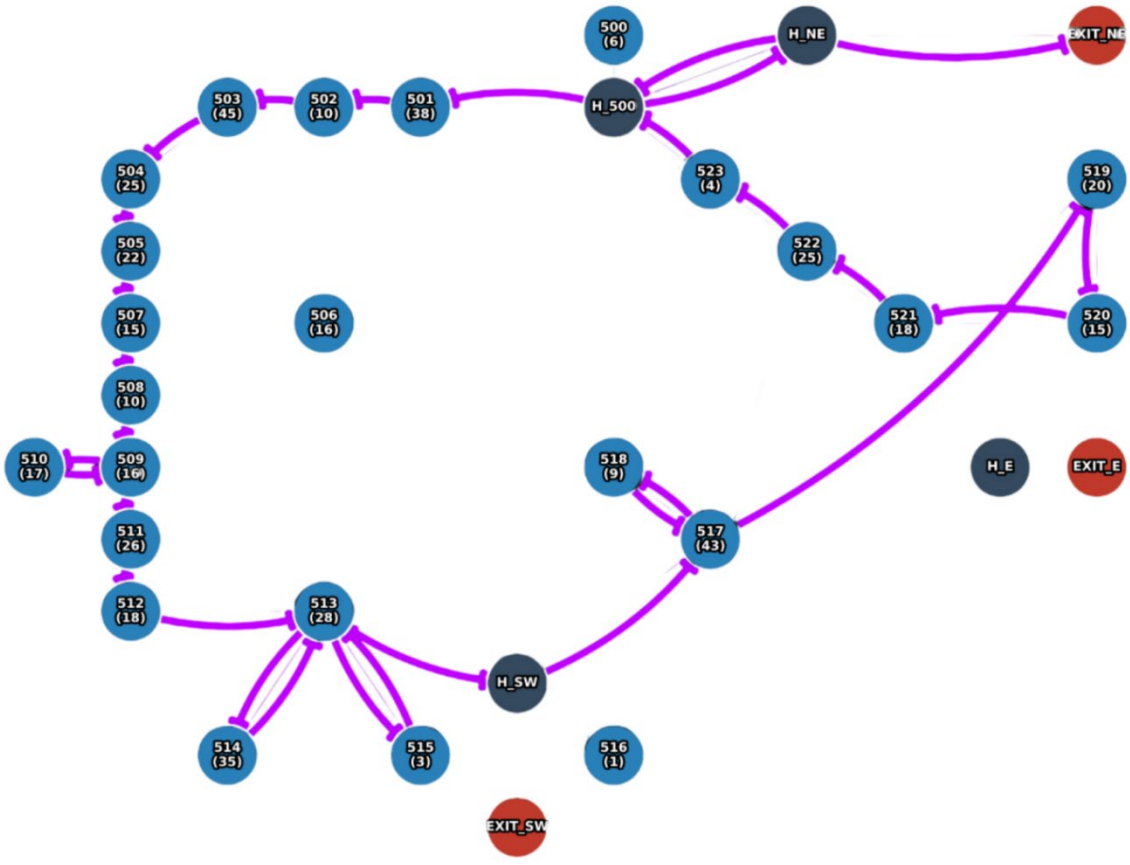


Figure 4. A* path at budget 30 min (Start H_NE, TER = 455).

4. Validation on Synthetic Museum Layout

To assess generalizability, algorithms were tested on the synthetic Metropoli Gallery, comprising 28 nodes (3 hallway hubs, 21 galleries, 3 exits, 1 main hallway), an average degree of 2.5, and 460 total exhibits. Results strongly parallel the MoMA findings: uninformed methods and Hybrid-ADFS plateau at 68 exhibits; Greedy achieves 233 exhibits with ~40 expansions; A* reaches full coverage (460 exhibits) at budgets ≥60 minutes with up to 99,000 expansions. These findings confirm that the observed trade-offs hold across different museum-like topologies with sparse connectivity.

Table 3. Algorithm performance on a synthetic museum-like layout (Metropoli Gallery).

Budget (min)	Algorithm	TER	Expansions	Path Length
10	BFS	68	28	6
	DFS	68	31	6
	Greedy	68	31	6
	A*	136	883	10
	Hybrid-ADFS	68	48	6
30	BFS	68	28	6
	DFS	68	36	6
	Greedy	233	40	14
	A*	456	67,927	29
	Hybrid-ADFS	68	41	6
60	BFS	68	28	6
	DFS	68	36	6
	Greedy	233	40	14
	A*	460	99,356	31
	Hybrid-ADFS	68	41	6
90	BFS	68	28	6
	DFS	68	36	6
	Greedy	233	40	14
	A*	460	99,356	31
	Hybrid-ADFS	68	41	6

5. Crowd-Dependent and Time-Decaying Rewards

A realistic reward model was applied, introducing: (1) a 40% time-of-day crowding penalty during peak hours (11:00–15:00), 15% otherwise; (2) progressive visitor fatigue reducing rewards up to 50% as more exhibits are collected (3% per item); (3) a 50% higher penalty in high-attendance galleries. Rewards are never reduced below 20% of base value. Under dynamic constraints, the performance gap narrows dramatically. At medium budgets (30 min), DFS achieves the highest reward (230) by committing early to deep high-reward chains before fatigue and peak penalties accumulate. At longer budgets (60 min), Greedy reaches essentially the same reward ceiling as A* (242 vs. 243) despite orders-of-magnitude lower expansions. These results contradict the assumption that more sophisticated always means better.

Table 4. Performance under crowd-dependent rewards - MoMA layout (averaged across starts)

Budget (min)	Algorithm	TER	Expansions	Path Length
10	BFS	85	132	9
	DFS	131	460	10
	Greedy	131	143	10
	A*	131	428	10
30	BFS	85	751	9
	DFS	230	8,826	29
	Greedy	229	2,449	30
	A*	229	10,527	30
60	BFS	85	1,681	9
	DFS	238	32,798	45
	Greedy	242	7,822	48
	A*	243	44,883	47
90	BFS	85	2,611	9
	DFS	238	58,358	45
	Greedy	242	12,472	48
	A*	243	81,303	47

6. Algorithm Selection Guidelines

Based on experimental results, practical guidelines for algorithm selection: (1) Very short visit (B < 10 min): Greedy or A*, both achieve maximum feasible reward quickly. (2) Typical visit (10 < B < 30 min): Greedy, highest practical reward (199 exhibits) with constant low expansions (~52 nodes). (3) Extended visit (30 < B < 60 min): A*, reaches near-complete or full coverage (455–465 exhibits), where Greedy plateaus. (4) Real-time mobile app: Greedy, consistent, fast execution, low memory, strong reward. (5) Offline planning: A*guarantees an optimal solution despite higher resource use. (6) Hybrid-ADFS: not recommended in current form; did not improve reward or coverage over baselines.

VI. DISCUSSION

1. Comparison with Previous Approaches

Unlike previous museum-routing and spatial path-planning studies [1], [4], [5], [6], this work evaluates five algorithms across multiple budgets and constraints on a real-world layout. The reward-constrained formulation highlights trade-offs between exhibit exposure and computational cost. Hybrid-ADFS extends prior heuristic combinations [11], [12] but requires further parameter tuning.

2. Practical Implementation Guidelines

Greedy Best-First Search is recommended for real-time mobile applications because it provides a strong reward with low computational cost (199 exhibits, 52 expansions, <0.01 s). A* is preferable for offline planning or extended visits when maximum exposure is prioritized (up to 465 exhibits, 2.38M expansions, >1 GB memory). BFS/DFS are primarily suitable for minimal tours, whereas Hybrid-ADFS is not recommended in its current

form. In museum design, graph connectivity appears more influential than reward distribution, suggesting that optimizing spatial layout may yield greater benefits than fine-grained exhibit weighting.

3. Limitations and Future Research Directions

The study has six main limitations: (1) a single museum case study, which may limit generalizability; (2) static exhibit rewards; (3) deterministic unit travel times; (4) no visitor personalization; (5) a limited budget range of up to two hours; and (6) no direct comparison with modern metaheuristic or learning-based approaches such as genetic algorithms, ant colony optimization, or reinforcement learning. These limitations restrict generalization and comparison with newer methods. Future work will include validation on additional museum layouts; integration of real-time crowd and time-varying rewards; comparison with genetic algorithms, ant colony optimization, and reinforcement learning [14], [15], [20]; multi-floor optimization; field studies with real visitors; and adaptive refinement of Hybrid-ADFS.

VII. CONCLUSION

This paper compared five search algorithms for reward-constrained museum visitor routing on a real-world model of the MoMA 5th floor. A* achieved complete coverage (465 exhibits at budgets ≥60 min) but required substantially greater computational resources, while Greedy provided strong practical performance with low computational cost. BFS and DFS consistently achieved lower exposure. Statistical analysis confirmed significant differences between A* and Greedy at shorter budgets, while their performance converged under realistic crowd-dependent rewards. These results demonstrate that simple greedy strategies can approach the performance of more computationally expensive methods under practical visitor constraints. The methodology can also be applied to other spatial routing problems with discrete rewards and resource constraints.

REFERENCES

[1]. Choi, K.C., et al. Genetic algorithm for tourism route planning considering time constraints—International Journal of Engineering Trends and Technology (2022).
[2]. Hsieh, Y.C., You, P.S. An efficient encoding scheme for a new multiple-type museum visitor routing problem—International Journal of Computational Intelligence Systems (2017).
[3]. Luo, X.G., et al. Trip planning for visitors in a service system with capacity constraints. Computers & Operations Research (2022).
[4]. Schmitt, S., et al. Fast routing graph extraction from floor plans. ResearchGate (2017).
[5]. Lin, S.W., Yu, V.F., Chou, S.Y. The museum visitor routing problem. Applied Mathematics and Computation (2010).
[6]. Sukhbaatar, U. Heuristic approach to solving the time-dependent team orienteering problem with multiple constraints. Proceedings EASTS (2021).
[7]. Vansteenwegen, P., Souffriau, W., Van Oudheusden, D. The orienteering problem: A survey. European Journal of Operational Research (2011).
[8]. Gunawan, A., Lau, H.C., Vansteenwegen, P. Orienteering problem: A survey of recent variants, solution approaches and applications—European Journal of Operational Research (2016).
[9]. Vansteenwegen, P., et al. Iterated local search for the team orienteering problem with time windows. Computers & Operations Research (2009).
[10]. Gavalas, D., et al. Heuristics for the time-dependent team orienteering problem. Computers & Operations Research (2015).

[11]. Souffriau, W., et al. A guided local search metaheuristic for the team orienteering problem. European Journal of Operational Research (2009).
[12]. Dang, D.-C., et al. An effective PSO-inspired algorithm for the team orienteering problem. European Journal of Operational Research (2012).
[13]. Thrun, S., et al. Probabilistic algorithms and the interactive museum tour-guide robot Minerva. International Journal of Robotics Research (2000).
[14]. Yadav, A.K., Sharma, P., Cheng, X., Shukla, S. Reinforcement Learning-Based Intelligent Path Planning for Optimal Navigation in Dynamic Environments. Neural Processing Letters, Springer, 2026
[15]. Rosa, A.C., et al. EvacuAI: An Analysis of Escape Routes in Indoor Environments with the Aid of Reinforcement Learning. Sensors, 23(21), 8892, MDPI, 2023.
[16]. Halder, S., Lim, K.H., Chan, J., Zhang, X. A survey on personalized itinerary recommendation: From optimization to deep learning. Applied Soft Computing, Vol. 152, 2023.
[17]. Hart, P.E., Nilsson, N.J., Raphael, B. A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics (1968).
[18]. Cormen, T.H., et al. Introduction to Algorithms, 3rd ed. MIT Press (2009).
[19]. Pearl, J. Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley (1984).
[20]. Juan, A.A., Marugan, C.A., Ahsini, Y., et al. Using Reinforcement Learning to Solve a Dynamic Orienteering Problem with Random Rewards. Batteries, 9(8), 416, MDPI, 2023.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Cybersecurity Awareness and Threat Perception Constructs

Katarina Kostelić
Faculty of Informatics
Juraj Dobrila University of Pula
Pula, Croatia
katarina.kostelic@unipu.hr
ORCID: 0000-0002-7107-1696

Dario Kukuljan
E-learning Support Department
Juraj Dobrila University of Pula
Pula, Croatia
dario.kukuljan@unipu.hr

Mieta Bobanović Dasko
Faculty of Informatics
Juraj Dobrila University of Pula
Pula, Croatia
mieta.bobanovic.dasko@unipu.hr
ORCID:0000-0002-4917-9158

Abstract
This study maps how the constructs of cybersecurity awareness and threat perception are operationalized in survey-based empirical research. Using Web of Science and Scopus records, we analyze the thematic evolution of a survey-focused corpus and conduct a construct-level analysis of screened studies by coding contexts, threat types, and extracted constructs. We then examine their interrelations through bipartite and construct-similarity networks. The findings show a shift from general cybersecurity and information security awareness toward more specific models of phishing, privacy, data breaches, identity theft, online scams, and social engineering. Across these contexts, perceived vulnerability, perceived severity, self-efficacy, response efficacy, knowledge, attitude, trust, privacy concern, and protective behavior form a recurring conceptual core. However, similar constructs appear under different labels and are applied inconsistently across user groups and threat scenarios. The paper provides a structured overview of the construct space and highlights the need to more clearly distinguish among awareness, knowledge, perception, efficacy, and behavior.

Keywords: cybersecurity awareness, threat perception, perceived vulnerability, perceived severity, survey research, construct mapping, phishing, privacy

I. INTRODUCTION

Digital threats increasingly depend on how users perceive, interpret, and respond to risk. Phishing, online scams, identity theft, data breaches, privacy invasion, ransomware, and social engineering all involve a human-facing dimension in which awareness, knowledge, perceived vulnerability, perceived severity, trust, self-efficacy, and attitude may shape behaviour and outcomes [1], [2], [3], [4], [5], [6], [7], [8]. As a result, cybersecurity awareness has become a recurring topic in empirical cybersecurity, information systems management, privacy, and human-computer interaction research. Cybersecurity awareness is often used broadly, but this paper does not treat it as a synonym for information, knowledge, or behavior. Following the situation-awareness tradition, awareness is understood as situated recognition and interpretation of threat-relevant cues: users must notice elements of a cybersecurity situation, comprehend their meaning, and relate them to possible future consequences or responses [9], [10], [11]. This distinction is important because information or knowledge alone may be insufficient to build awareness that leads to an effective response [12], [13], [14]. Empirical research on situational and cybersecurity awareness can be broadly divided into studies that measure situational or cybersecurity awareness using task-based, behavioral, or performance-oriented indicators, and studies that examine self-reported awareness, perceptions, appraisals, and behavioral constructs via questionnaires. This paper focuses on the latter strand. This self-reported strand includes studies that draw on awareness scales, protection motivation theory, technology threat avoidance theory, privacy concern, and behavioral intention models [1], [4], [5], [6], [15], [16]. However, the construct space remains fragmented. Different studies use partially overlapping terms such as cybersecurity awareness, information security awareness, privacy awareness, threat perception, risk perception, perceived vulnerability, perceived susceptibility, perceived severity, privacy concern, self-efficacy, response efficacy, trust, attitude, intention, and behavior. These constructs are often embedded in specific

empirical settings, such as phishing, mobile identity protection, social media use, online banking, health-tracker privacy, employee training, or student awareness. Although these studies contribute valuable local findings, it is difficult to see how the constructs relate to one another across the wider field.

This paper addresses this gap by mapping the constructs of cybersecurity awareness and threat perception in survey-based research. The focus is deliberately narrowed to empirical survey studies in which these constructs are measured, modeled, or used to explain user responses to cyber threats. The analysis, therefore, treats the field as a set of recurring concepts, empirical contexts, threat types, and relationships. The following research questions guide the study:

- RQ1. How has the thematic structure of survey-based research on cybersecurity awareness and threat perception evolved?
- RQ2. Which contexts, threat types, and constructs recur most frequently in the screened studies?
- RQ3. How are contexts, threats, and constructs interrelated?
- RQ4. Which construct communities can be identified in the construct-similarity network?

The contribution of the paper is twofold. First, it provides a structured map of the construct vocabulary used in survey-based research on cybersecurity awareness and threat perception. Second, it identifies fragmentation and overlap in the use of constructs, supporting future work on clearer conceptual and measurement alignment.

II. METHOD

1. Data Collection and Preparation

The analysis was based on bibliographic records retrieved from Web of Science and Scopus. The search was conducted in Web of Science and Scopus using awareness- and threat-perception-related terms combined with cyber/digital context terms and threat examples: TITLE-ABS-KEY((“threat awareness” OR “vulnerability awareness” OR “cybersecurity awareness” OR “information security awareness” OR “privacy awareness” OR “threat perception” OR “vulnerability perception” OR “perceived vulnerability” OR “perceived severity” OR “perceived threat”) AND (cyber* OR digital OR online OR internet) AND (phishing OR scam* OR “identity theft” OR ransom* OR impersonation OR “account compromise” OR “account hijacking” OR “unauthorized access” OR “data breach” OR “data misuse” OR “data exposure” OR “privacy invasion” OR “unauthorized sharing” OR “personal data”) AND (survey OR questionnaire OR scale OR instrument OR “measurement model” OR “self-report” OR “structural equation model*” OR SEM OR PLS-SEM OR “partial least squares”) AND NOT (“raising awareness” OR “raise awareness” OR “increase awareness” OR “improve awareness” OR “awareness campaign*”)). The search query, preprocessing, descriptive results, thematic maps, descriptions, and screening and network outputs are provided in the Supplementary file. The Scopus query returned 174 documents, while the corresponding Web of Science Core Collection query returned 93 documents. The deduplicated corpus covered the period from 2004 to 2026 and included 194 documents published across 166 sources. The dataset is relatively recent, with an average document age of 3.82 years, and shows continued growth at an annual rate of 11.96%. The records contained 737 Keywords Plus and 681 author keywords, indicating a sufficiently rich vocabulary for thematic and construct-oriented analysis. A focused screening stage followed the bibliometric dataset. Abstracts and keywords were screened to identify studies relevant to survey-based research on cybersecurity awareness and threat perception constructs. Each record was coded for inclusion status, reason for exclusion where applicable, study type, empirical context, threat type, constructs identified, measured dimensions or indicators (if applicable), and availability of questionnaire items (based on the statement in the abstract). Studies were included when the record indicated an empirical survey, scenario-based survey,

questionnaire-based model, or survey-supported empirical analysis involving cybersecurity awareness, information security awareness, privacy awareness, perceived vulnerability, perceived severity, perceived threat, risk perception, efficacy, trust, concern, attitude, intention, behavior, or related user-facing constructs. Studies were excluded if they were primarily technical, conceptual, review-based, training- or taxonomy-oriented, or not focused on measured user constructs.

For retained studies (n = 112), we coded the available empirical context, threat type, and construct labels in abstracts. Context refers to the user group, setting, or application domain. Threat type refers to the cyber- or privacy-related risk under study. For this study, constructs were defined as named psychological, perceptual, attitudinal, motivational, knowledge-related, or behavioral concepts that were measured, modeled, or explicitly treated as empirical variables in survey-based studies. Broader technical topics, intervention types, system features, and general keywords were not treated as constructs unless they were operationalized as measured user-facing variables.

2. Bibliometric and Network Analysis

The first part of the analysis examined the survey-oriented corpus's main information and thematic evolution [17], [18], [19]. This was used to identify how the research field changed over time and which topics became more central in later periods. For this purpose, the bibliometric analysis provided a temporal and thematic overview of the corpus. The second stage focused on screening studies and extracting contexts, threats, and constructs. Frequency summaries were created based on their occurrence across studies. Bipartite networks linking context–threat, context–construct, and threat–construct pairs that appeared in the same screened studies; edge weights indicate the number of studies in which each pair co-occurred. To map terminological and conceptual overlap in the construct vocabulary, the construct-similarity network was created from cleaned construct labels. Nodes represent constructs, and edges connect construct pairs that share at least two meaningful terms after normalization; edge weights indicate the number of shared terms. Community detection was applied to the giant component to identify clusters of related construct labels.

III. RESULTS

1. Thematic Evolution

The thematic evolution map (Supplementary file) indicates a gradual consolidation of the survey-focused corpus around privacy, information security, cybersecurity, and survey-based measurement (Table 1; visual representations are available in the Figures and interactive graphs sections of the Supplementary files). The thematic evolution does not indicate a simple sequence in which earlier themes disappear and new themes replace them. Instead, Table 1 shows a cumulative narrowing and reorientation of the field. The broad theme of information dissemination represents the earliest period, while the 2013–2020 period adds trust, surveys, data privacy, and data security. In the 2021–2026 period, several of these themes remain visible, especially privacy, data privacy, information security, and surveys, but they appear alongside cybersecurity and motivation. The overlap across periods is therefore informative: it suggests continuity in the privacy and information-security vocabulary, while the later appearance of cybersecurity and motivation indicates a stronger focus on user-facing empirical models of threat perception, awareness, and response. Thus, the main trend is not thematic replacement, but increasing specification. The survey-based research moves from general information-oriented concerns to a survey-based examination of how users understand privacy- and security-related threats and what motivates their responses. Because the thematic-evolution results remain broad and partly overlapping, the following construct-level analysis examines more directly how these themes are operationalized through empirical contexts, threat types, and measured constructs.

Table 1. Topics in Clusters derived by Thematic Evolution

	Period	Dominant themes
1.	2004–2010	Information dissemination
2.	2013 – 2020	Trust; surveys; data privacy; security of data; privacy
3.	2021 – 2026	Cybersecurity; privacy; information security; data privacy; motivation; surveys

2. Contexts, Threats, and Constructs

The screened studies show a heterogeneous empirical landscape. Table 2 reports independent frequency lists for the screened studies, where n denotes the number of papers in which a given category was coded; categories appearing in the same row are not necessarily related within the same papers. The most frequent contexts are users, students, universities, employees, and social or online settings, while threat types are more concentrated around phishing, which appears in 24 papers. The construct vocabulary is broader, but its recurring core comprises cybersecurity (18 papers) or information security awareness (15 papers), along with perceived severity, perceived vulnerability, self-efficacy, privacy awareness, privacy concerns, attitude, perceived threat, and response efficacy. Because Table 2 reports only separate frequency rankings, the following bipartite networks examine how contexts, threats, and constructs co-occur.

Table 2. Frequent Contexts, Threats, and Constructs

	Contexts, types, and constructs - independent ranked frequency lists; rows do not indicate co-occurrence					
	Context type	***n papers***	***Threat type***	***n papers***	***Construct type***	***n papers***
1.	User	36	Phishing	24	Cybersecurity awareness	18
2.	Student	23	Data breaches	10	Perceived severity	16
3.	University	16	Identity theft	8	Perceived vulnerability	16
4.	Employee	15	Cyberattacks	7	Information security awareness	15
5.	Social	12	Social engineering	5	Self-efficacy	14
6.	Medium	9	Privacy invasion	5	Privacy awareness	12
7.	Online	9	Cybercrime	5	Privacy concerns	11
8.	Context-specific	8	Scams	4	Attitude	7
9.	Indonesia	8	Online scams	4	Perceived threat	6

	Contexts, types, and constructs - independent ranked frequency lists; rows do not indicate co-occurrence					
	Context type	***n papers***	***Threat type***	***n papers***	***Construct type***	***n papers***
10.	Phishing context	8	Cyber threats	4	Response efficacy	6

3. Interrelations among Contexts, Threats, and Constructs

The threat–construct bipartite network (Table 3; visual representation available in the Supplementary files) is organized around a few central bridges rather than evenly distributed links. The strongest edges link phishing with cybersecurity awareness and information security awareness, while data breaches and phishing attacks also connect to cybersecurity awareness. Privacy invasion appears as a separate privacy-related link, based on perceived severity. Overall, the network suggests a phishing-centered awareness core, with privacy-related threats forming a smaller but distinct strand.

The context–construct network (Table 3; visual representation available in the Supplementary files) shows that construct use is distributed across both general and more specialized empirical settings. The strongest edges connect cybersecurity awareness with Malaysia and university students, while fitness tracker users connect to attitude toward technology, coping behavior, and privacy concerns. This suggests that awareness constructs are prominent in broader educational or national contexts, whereas specialized technology-use contexts are associated with narrower privacy- and coping-related constructs.

The context–threat network (Table 3; visual representation available in the Supplementary files) confirms that phishing is the main bridge between empirical settings and threat types. The strongest edges link phishing to employees, university students, and Australia; malware to employees, and unauthorized access to Saudi Arabia. This indicates a visible phishing-centered core, while other threats appear in more context-specific links. The bipartite networks show how contexts, threats, and constructs co-occur; the next analysis focuses on communities within the construct vocabulary itself.

Table 3. Top Edges in Bipartite Networks

	Bipartite network	**Node 1**	**Node 2**	**Edge weight (n)**
1.	Threat–construct	phishing	cybersecurity awareness	6
2.	Threat–construct	phishing	information security awareness	4
3.	Threat–construct	data breaches	cybersecurity awareness	3
4.	Threat–construct	phishing attacks	cybersecurity awareness	3
5.	Threat–construct	privacy invasion	perceived severity	3
1.	Context–construct	malaysia	cybersecurity awareness	3
2.	Context–construct	university students	cybersecurity awareness	3
3.	Context–construct	fitness tracker users	attitude toward technology	2
4.	Context–construct	fitness tracker users	coping behavior	2

	Bipartite network	Node 1	Node 2	Edge weight (n)
5.	Context–construct	fitness tracker users	privacy concerns	2
1.	Context–threat	employees	phishing	3
2.	Context–threat	university students	phishing	3
3.	Context–threat	australia	phishing	2
4.	Context–threat	employees	malware	2
5.	Context–threat	saudi arabia	unauthorized access	2

4. Construct Communities in the Giant Component of the Construct-Similarity Network

The giant component of the construct-similarity network (Figure 1) shows that the largest communities center on information security awareness and behavior, data privacy literacy and privacy-loss concerns, cybersecurity awareness and risk perception, personal information disclosure, perceived risk and control, privacy protection behavior, privacy awareness and concerns, and online data misuse. These communities suggest that empirical assessments of cybersecurity awareness rarely treat awareness as a narrow, standalone construct. Instead, a substantial part of what is measured as, or around, awareness falls on perception and appraisal: perceived risk, perceived vulnerability, perceived severity, perceived control, perceived behavioral control, perceived data sensitivity, and related concern-based constructs. At the same time, these appraisal-oriented constructs are not uniformly organized across the field but appear in contextual communities shaped by vocabularies of privacy, data, information security, cybersecurity, literacy, disclosure, and protection.

For example, information security awareness is closely connected to information security behavior and organisationally oriented security constructs. In contrast, data privacy literacy is linked to fears of privacy loss and health data privacy concerns. Cybersecurity awareness is linked to risk perception and value of privacy, whereas privacy awareness and privacy concerns form a smaller but conceptually coherent privacy-focused cluster. This structure indicates that awareness in the broader empirical literature is often operationalized through a situated understanding of what is at risk, how serious the threat is, how personally relevant it is, and what control or response options are perceived as available.

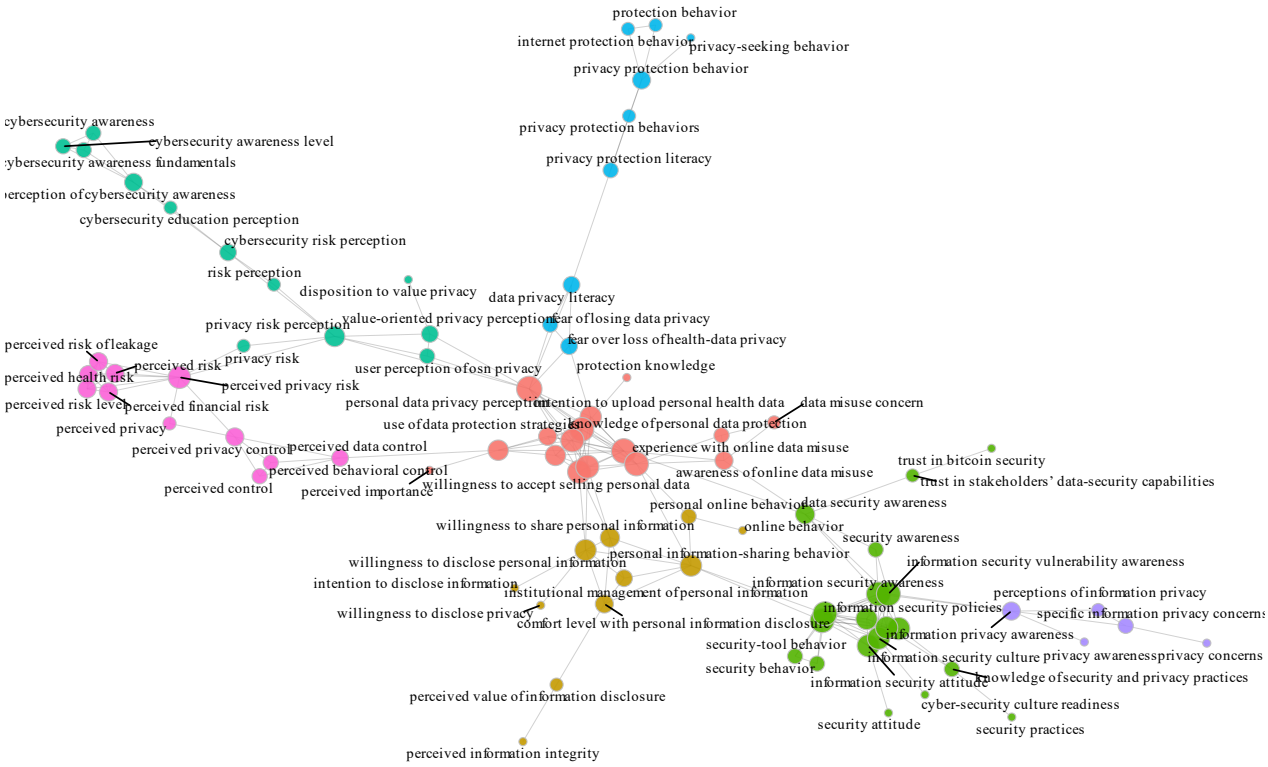


Figure 1. Construct-similarity network of the giant component

Note: Node size represents degree; colors indicate detected communities; edges connect construct labels sharing at least two meaningful terms after preprocessing. The layout is used for visual interpretation of construct proximity, not causal ordering. An interactive version is available in the Supplementary files.

IV. DISCUSSION AND CONCLUSION

The results answer the four research questions by showing that survey-based cybersecurity awareness and threat perception research has become more specific, more empirically measurement-oriented, and more heterogeneous at the construct level. The thematic evolution (RQ1) indicates a shift from broader information dissemination and information-security themes toward privacy, data privacy, cybersecurity, information security, surveys, and motivation. This suggests that the field has moved from general awareness and information-security concerns toward more specific empirical models of how users perceive, appraise, and respond to digital threats. Regarding the frequency of contexts, threat types, and constructs (RQ2), the screened studies show that the literature is concentrated on several recurring yet unevenly represented topics. Users, students, universities, employees, social media, online settings, and country-specific samples dominate the empirical contexts, while phishing is the most frequent threat type. Other threats, including data breaches, identity theft, social engineering, privacy invasion, cybercrime, and scams, appear less frequently and are more dispersed. At the construct level, the core constructs consist of cybersecurity awareness, information security awareness, perceived severity, perceived vulnerability, self-efficacy, privacy awareness, privacy concerns, attitude, perceived threat, and response efficacy. This pattern shows that awareness research is rarely limited to awareness alone but usually combines it with threat appraisal, efficacy appraisal, concern, attitude, and behavioral response. Regarding the relationships among contexts, threats, and constructs (RQ3), the bipartite

networks exhibit a sparse structure. Phishing serves as the clearest bridge among threat types, user contexts, and awareness-related constructs, especially in studies of employees and university students. Privacy invasion, data breaches, identity theft, malware, and unauthorized access are connected to narrower, more context-specific parts of the corpus. These results suggest that the literature has a phishing-centered empirical core, whereas other cyber and privacy threats are less prevalent across contexts and combinations of constructs.

Regarding the construction of communities in the construct-similarity network (RQ4), the giant component is organized into several construct families related to information and knowledge, cybersecurity and privacy awareness, threat and risk appraisal, efficacy appraisal, trust and concern, and behavioral response. This heterogeneity is not only terminological. It reflects different assumptions about where awareness is located in the process that leads from cyber-related information to protective or risky behavior. Based on these findings, we propose an awareness–appraisal–response framework (Figure 2). The framework starts with the digital context and the threat object because awareness is always situated: phishing, scams, data breaches, privacy invasion, and identity theft provide different cues and require different responses. Information and knowledge represent the factual and conceptual resources users may draw on to achieve awareness. In the narrow sense, cybersecurity awareness refers to situated recognition of threat-relevant cues. In the broader sense, awareness also involves comprehension and projection: users must interpret the cues, appraise their implications, and relate the situation to possible consequences and responses. In this sense, awareness functions as an interpretive bridge between general knowledge and later judgments or actions.

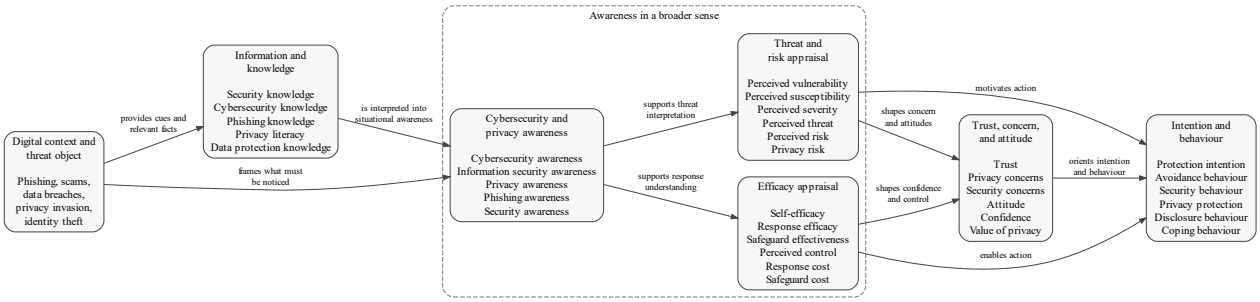


Figure 2. Integrative framework of the constructs in the research on cybersecurity awareness

Once a situation is recognized as security-relevant, the framework treats threat appraisal and efficacy appraisal as explicit construct domains that cover vulnerability, susceptibility, severity, threat/risk, self-efficacy, response efficacy, perceived control, safeguard effectiveness, and safeguard cost. These domains explain why awareness does not automatically lead to protective behavior: users may recognize a threat but feel unable to respond, distrust the recommended action, or judge the action as too costly. Trust, concern, and attitude then shape orientation toward action. The framework positions intention and behavior as outcomes of this process, including protection intention, avoidance behavior, security behavior, privacy protection, disclosure, and coping behavior. This positioning is consistent with studies in which phishing-related threat perception, perceived severity, perceived susceptibility, self-efficacy, safeguard effectiveness, and safeguard cost are linked to avoidance behavior [4], and with identity-protection research that distinguishes perceived threat appraisal, self-efficacy, motivational intention, awareness, and actual identity-protection actions [16]. Privacy-oriented studies show a similar pattern: privacy awareness, privacy risk, privacy concerns, privacy invasion experience, and perceived control are linked to self-disclosure [6], while privacy concern and perceived threat shape post-violation coping behavior and attitude [15]. Recent studies on mobile identity protection and online blackmail further show that awareness, efficacy, attitude, motivation, and behavior may occupy different positions within empirical models rather than function as interchangeable indicators [3], [20]. These studies jointly support the view that knowledge, awareness, appraisal, efficacy, attitude, intention, and behavior are interlinked, but should not be treated as interchangeable indicators. This structure highlights why treating cybersecurity awareness as a

synonym for knowledge or behavior is conceptually problematic. Information may be available, and knowledge may exist, without necessarily producing cue recognition, threat interpretation, response confidence, or protective action. The proposed framework is not a tested causal model, but a synthesis of the construct families identified in the mapping. Its practical implication is that awareness interventions should not stop at information provision but should also support cue recognition, threat interpretation, response confidence, and the translation of responses into protective or avoidance behavior. The study has several limitations. The analysis depends on the selected search query, the bibliographic database coverage, and the extraction of constructs from the keywords and abstracts of screened records, so some relevant studies and alternative construct labels may have been missed. The network analyses capture co-occurrence among coded elements and similarity among construct labels, not validated theoretical or causal relationships between constructs. Future research should therefore examine measurement items in greater detail, compare how similar constructs are operationalized across studies, and test the proposed framework empirically in specific threat contexts such as phishing, scams, data breaches, privacy invasions, or identity theft. A useful next step would be to develop a more coherent measurement model that distinguishes among information, knowledge, awareness, appraisal, efficacy, concern, and behavior, while allowing comparison across user groups and threat scenarios.

SUPPLEMENTARY FILES

Supplementary materials (Supplementary file-analysis, Figures, interactive graphs, and Source files) are available at: https://doi.org/10.17605/OSF.IO/789JR.

ACKNOWLEDGEMENT

The EU NextGeneration funded this work under the institutional research project at Juraj Dobrila University of Pula, number IIP_UNIPU_010133. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Juraj Dobrila University of Pula.

REFERENCES

[1]. [1] A. McCormac et al., “Test-retest reliability and internal consistency of the Human Aspects of Information Security Questionnaire (HAIS-Q),” 2016.

[2]. [2] C. Blackwood-Brown, Y. Levy, and J. D’Arcy, “Cybersecurity awareness and skills of senior citizens: a motivation perspective,” *J. Comput. Inf. Syst.*, vol. 61, no. 3, pp. 195–206, May 2021, doi: 10.1080/08874417.2019.1579076.

[3]. [3] B. AlGhanboosi, S. Ali, and A. Tarhini, “Examining the effect of regulatory factors on avoiding online blackmail threats on social media: A structural equation modeling approach,” *Comput. Hum. Behav.*, vol. 144, p. 107702, Jul. 2023, doi: 10.1016/j.chb.2023.107702.

[4]. [4] N. A. G. Arachchilage, S. Love, and K. Beznosov, “Phishing threat avoidance behavior: An empirical investigation,” *Comput. Hum. Behav.*, vol. 60, pp. 185–197, Jul. 2016, doi: 10.1016/j.chb.2016.02.065.

[5]. [5] Z. Tang, A. S. Miller, Z. Zhou, and M. Warkentin, “Does government social media promote users' information security behavior towards COVID-19 scams? Cultivation effects and protective motivations,” *Gov. Inf. Q.*, vol. 38, no. 2, p. 101572, Apr. 2021, doi: 10.1016/j.giq.2021.101572.

[6]. [6] G. O. A. Ampong et al., “Examining self-disclosure on social networking sites: a flow theory and privacy perspective,” *Behav. Sci.*, vol. 8, no. 6, p. 58, Jun. 2018, doi: 10.3390/bs8060058.

[7]. [7] T. I. Supti *et al.*, “Digital dependency and security risk: investigating the connections between fear of missing out, problematic social media use, and vulnerability perceptions,” *Journal of Technology in Behavioral Science*. Springer, 2025. doi: 10.1007/s41347-025-00515-0.

[8]. [8] R. J. Mejias, J. J. Greer, G. C. Greer, M. M. Shepherd, and R. Y. Reyes, “A model for information security vulnerability awareness,” *Computers and Security*, vol. 151. Elsevier Ltd, 2025. doi: 10.1016/j.cose.2024.104305.

[9]. [9] M. R. Endsley, “Measurement of situation awareness in dynamic systems,” *Hum Factors*, vol. 37, no. 1, pp. 65–84, Mar. 1995, doi: 10.1518/001872095779049499.

[10]. [10] H. J. Ofte and S. Katsikas, “Understanding situation awareness in SOCs, a systematic literature review,” *Computers & Security*, vol. 126, p. 103069, Mar. 2023, doi: 10.1016/j.cose.2022.103069.

[11]. [11] J. Li and A. Y. K. Chua, “Think or respond: understanding the impact of cognitive appraisals on threat detection and phishing susceptibility,” *Proceedings of the Association for Information Science and Technology*, vol. 62, no. 1. John Wiley and Sons Inc, pp. 384–395, 2025. doi: 10.1002/pra2.1264.

[12]. [12] E. Blasch, D. Shen, K. D. Pham, and G. Chen, “Review of game theory applications for situation awareness,” in *Sensors and Systems for Space Applications VIII*, SPIE, 2015, pp. 141–150.

[13]. [13] K. Kostelic, “Guessing the game: an individual’s awareness and assessment of a game’s existence,” *Games*, vol. 11, no. 2, p. 17, Mar. 2020, doi: 10.3390/g11020017.

[14]. [14] K. Kostelić, “Dynamic awareness and strategic adaptation in cybersecurity: a game-theory approach,” *Games*, vol. 15, no. 2, p. 13, Apr. 2024, doi: 10.3390/g15020013.

[15]. [15] K. Baskaran and S. K. Mathew, “Understanding coping intentions of fitness tracker users: an empirical investigation using fear appeals,” *Int. J. Hum.-Comput. Interact.*, vol. 40, no. 3, pp. 795–807, Feb. 2024, doi: 10.1080/10447318.2022.2124358.

[16]. [16] K. S. Jones, E. K. Foster, M. E. Armstrong, and A. Siami Namin, “Do users’ mental models of hackers affect their perceived susceptibility to and the perceived severity of phishing attacks?” *Inf. Comput. Secur.*, Feb. 2026, doi: 10.1108/ICS-04-2023-0052.

[17]. [17] V. Batagelj and M. Cerinšek, “On bibliographic networks,” *Scientometrics*, vol. 96, no. 3, pp. 845–864, 2013.

[18]. [18] M. Aria and C. Cuccurullo, “Bibliometrix: An R-tool for comprehensive science mapping analysis,” Journal of Informetrics, vol. 11, no. 4, pp. 959–975, 2017.

[19]. [19] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, “An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field,” *Journal of Informetrics*, vol. 5, no. 1, pp. 146–166, 2011.

[20]. [20] Y. Alhelaly, G. Dhillon, and T. Oliveira, “Mobile identity protection: the moderating role of self-efficacy,” *Australasian Journal of Information Systems*, vol. 28. Australasian Association for Information Systems, pp. 1–40, 2024. doi: 10.3127/ajis.v28.4397.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Model Predictive Control-Based Trajectory Tracking and Dynamic Obstacle Avoidance for an Autonomous Vehicle in Robot Operating System 2

İbrahim Toy
R&D Center
MPG Inc.
Konya, Türkiye
itoy@mpg.com.tr
ORCID:0000-0002-7365-9661

Akif Durdu
Department of Electrical and Electronics Engineering
Konya Technical University
Konya, Türkiye
adurdu@ktun.edu.tr
ORCID:0000-0002-5611-2322

Abstract
Autonomous vehicles consist of layered modules including perception, sensor fusion, localization, motion planning, and control. Within this structure, the control module plays a critical role in accurate trajectory tracking and safe driving. Traditional path-tracking controllers such as Proportional-Integral-Derivative (PID), Pure Pursuit, and Stanley are widely used for their simplicity and low computational requirements. However, their performance may degrade in dynamic and high-speed scenarios. To address these limitations, a Model Predictive Control (MPC) based path tracking algorithm is implemented within Robot Operating System 2 (ROS 2) Humble framework on a physical autonomous golf vehicle equipped with Real-Time Kinematic Global Navigation Satellite System (RTK-GNSS), Inertial Measurement Unit (IMU), and Velodyne 3D Light Detection and Ranging (LiDAR) sensors. The MPC is formulated with a prediction horizon of Np = 20 and a control horizon of Nc = 10, and obstacle constraints derived from a real-time LiDAR occupancy grid are integrated directly into the optimization loop. Experimental validation on a 350 m closed-loop test route at speeds up to 7 m/s demonstrates a mean lateral error of 0.302 m and an RMS lateral error of 0.447 m, representing improvements of 41% and 38%, respectively, over a comparably tuned Stanley controller. Dynamic obstacle avoidance tests show a transient increase in lateral deviation of up to 0.48 m, confirming safe and stable navigation in mixed static and dynamic environments.

Keywords: autonomous vehicles, path tracking control, model predictive control, robot operating system 2 (ROS 2), trajectory tracking

I. INTRODUCTION

Autonomous ground vehicles rely on interconnected modules, including perception, sensor fusion, motion planning, decision-making, and control. Among these, a robust control architecture is critical for accurate and safe trajectory tracking [1]. Classical controllers such as Proportional-Integral-Derivative (PID), Stanley, and Pure Pursuit are effective in low-speed or structured environments, but their performance degrades in high-speed or complex scenarios. Advanced methods such as Linear Quadratic Regulator (LQR), Sliding Mode Control (SMC), and Model Predictive Control (MPC) enable better handling of system constraints, nonlinear dynamics, and uncertainties. MPC-based path-tracking methods use kinematic and dynamic vehicle models to optimize steering and speed while adapting to environmental uncertainties. Unlike classical controllers, MPC explicitly handles system constraints and provides real-time optimal control actions [2].

Control problems in autonomous driving are typically addressed using a hierarchical structure, in which a feasible trajectory is first generated and then tracked by a feedback controller. Although this approach was

successfully employed in the DARPA Grand Challenge competitions, it exhibits significant limitations near the limits of tire adhesion or during high-speed maneuvers. Consequently, optimal control techniques such as MPC, which jointly solve planning and control within a single optimization framework, have emerged as prominent alternatives [3].

In this study, an MPC-based path tracking controller has been implemented on an autonomous vehicle platform using the Robot Operating System 2 (ROS 2) framework. The controller optimizes steering and speed commands based on a reference trajectory, integrating sensor fusion outputs to enable obstacle avoidance. Although MPC-based trajectory tracking for autonomous vehicles has been addressed in numerous studies [1-3], the original contribution of this work is threefold: (1) a complete, real-time system integration of LiDAR-derived occupancy-grid obstacle constraints directly into the MPC optimization loop on a physical golf vehicle platform; (2) experimental validation under both static and dynamic obstacle scenarios at operational speeds of up to 7 m/s within the ROS 2 Humble framework; and (3) a systematic performance comparison against the Stanley controller under identical test conditions, demonstrating a 41% reduction in mean lateral error.

II. PATH TRACKING PROBLEM

One of the most critical topics in control is the path tracking problem [4]. To define the path-tracking problem for autonomous vehicles, a known reference trajectory is assumed. The objective is to ensure that the vehicle follows this path with minimum lateral and longitudinal errors. Control inputs (steering angle and speed) are computed to minimize deviations from the desired trajectory while complying with vehicle dynamics and actuator constraints [5].

Classical controllers such as PID, Stanley, and Pure Pursuit perform well at low speeds but may lose accuracy in high-speed or dynamic scenarios. In contrast, MPC predicts future vehicle behavior over a finite horizon and optimizes control actions accordingly [6]. This predictive capability enables more stable and accurate path tracking, particularly under challenging conditions involving nonlinear dynamics and varying path curvature.

In the present study, the vehicle operates at speeds of up to 7 m/s on a route with curve radii as tight as 8 m, yielding lateral accelerations that approach the limits at which purely reactive controllers such as Stanley begin to exhibit phase lag and overshoot. This operating regime motivates the use of MPC, whose predictive horizon allows the controller to begin corrective steering before the vehicle enters a high-curvature zone, unlike reactive controllers that respond only to the current measured error.

III. CONTROL METHODS

Autonomous vehicles employ various control strategies to achieve accurate path tracking. Classical controllers such as PID, Pure Pursuit, and the Stanley algorithm are simple and effective for low-speed or structured environments [7]. The PID controller adjusts steering based on lateral deviation, while the Stanley algorithm accounts for both lateral and heading errors to improve directional stability. However, these methods generally suffer from performance degradation in high-speed or dynamic scenarios. Optimal control approaches, including LQR, minimize tracking error while balancing control effort. Although computationally efficient, LQR cannot handle input constraints or predict future vehicle states [8]. Robust controllers such as SMC provide strong resistance to disturbances and uncertainties, guaranteeing stable path tracking even under challenging conditions. For trajectory representation, spline-based paths (e.g., B-spline or quintic spline) provide smooth and dynamically feasible reference trajectories, particularly suited for advanced controllers such as MPC. MPC can simultaneously optimize steering and speed over a prediction horizon, explicitly handle constraints, and improve

tracking accuracy in real time [9]. This work focuses on implementing and evaluating an MPC-based path-tracking controller within a ROS 2 framework and compares its performance with classical methods.

IV. MPC FUNDAMENTALS

MPC is an advanced optimal control strategy that predicts a system's future behavior and computes control actions over a finite prediction horizon. At each sampling step, only the first input is applied, and the optimization is repeated over a receding horizon. This predictive approach allows MPC to anticipate future vehicle states and handle system constraints such as steering limits, lateral acceleration, and tire slip, making it well-suited for safe and stable path tracking [10], [11]. The performance of MPC depends on the prediction horizon (Np), which determines how far ahead the system is predicted, and the control horizon (Nc), which defines how many control inputs are optimized. In this study, the prediction horizon was set to Np = 20 and the control horizon to Nc = 10, with a sampling period of 0.1 seconds. Three considerations guided the selection of Np = 20 and Nc = 10. First, the lookahead distance: with a sampling period of 0.1 s and a maximum speed of 7 m/s, Np = 20 provides a prediction horizon of 2 s, corresponding to a spatial look-ahead of up to 8 m — sufficient to encompass the tightest curve radii (~8 m) on the test route. Shorter horizons (Np < 10) were found in preliminary trials to produce oscillatory steering on high-curvature segments. Second, the control horizon ratio: setting Nc = Np/2 = 10 reduces the number of free optimization variables by half while retaining responsive control authority — a standard heuristic for balancing computation time and tracking performance [12]. Third, real-time feasibility: the chosen horizons resulted in a mean MPC solve time of 8.3 ms (max: 14.1 ms) on the DS-1400 embedded computer, comfortably within the 100 ms sampling period. Longer prediction horizons (Np = 30, Nc = 15) were evaluated, but increased the worst-case solve time to 22 ms without meaningful tracking improvement, so Np = 20 / Nc = 10 was selected as the operating point. Proper selection of these horizons balances tracking accuracy, stability, and computational efficiency [12].

The cost function in Equation (1) is central to MPC and penalizes deviations from the reference trajectory and abrupt changes in control inputs, thereby ensuring smooth and accurate tracking [13].

$$J = \sum_{i=1}^{N_p} \|y(k+i) - r(k+i)\|_Q^2 + \sum_{i=0}^{N_c-1} \|\Delta u(k+i)\|_R^2 \qquad (1)$$

In this formula:

- r_{k+i} denotes the reference output,
- y_{k+i} denotes the predicted system output,
- Δu_{k+i} denotes the control input increments,

Q and R are weighting coefficient matrices that penalize state tracking error and control effort, respectively. Np is the prediction horizon (Np = 20 in this study), and Nc is the control horizon (Nc = 10 in this study).
Vehicle dynamics are typically modeled using a bicycle model that captures lateral and yaw motions in a compact state-space representation [14]. The key vehicle parameters used in this study are: vehicle mass ≈ 450 kg, front-to-rear axle distance (wheelbase) of 1.85 m, maximum steering angle ±30°, maximum longitudinal speed of 7 m/s, and maximum longitudinal acceleration of 1.5 m/s². The MPC weighting matrices were tuned as Q = diag(10, 10, 5, 1) and R = diag(0.1; 0.1).
Several MPC variants exist:

- Classical MPC: Simple and computationally efficient for linear or mildly nonlinear systems.

- Adaptive MPC: Adjusts model parameters online to handle changing conditions such as slippery roads or varying vehicle mass.
- Robust MPC: Accounts for worst-case disturbances and modeling uncertainties to ensure safety in critical scenarios.
- Hybrid MPC: Combines MPC with other controllers (e.g., Stanley) to leverage complementary strengths for real-time applications [15].

In this study, a classical MPC framework was implemented in the ROS 2 environment to optimize steering and speed commands for autonomous vehicle path tracking, ensuring safe and accurate trajectory following even under dynamic conditions.

V. LIDAR-BASED PERCEPTION AND CONSTRAINTS INTEGRATION FOR OBSTACLE AWARE MPC

Safe autonomous driving requires accurate and real-time perception of the surrounding environment. In this study, a 3D LiDAR sensor is employed as the primary perception source. Raw LiDAR point cloud data is processed to construct a 2D occupancy grid map, which encodes the presence of static and dynamic obstacles in the vehicle's local environment. This occupancy grid is updated continuously during each control cycle and serves as the primary input to the obstacle-avoidance module. Individual sensors are inherently subject to noise, measurement uncertainty, and performance degradation under adverse environmental conditions. To mitigate these limitations, the perception pipeline incorporates IMU and Global Navigation Satellite System (GNSS) data alongside LiDAR measurements, enabling a more robust and consistent environmental representation. The fused sensor outputs are time-synchronized and fed into the MPC optimization loop with an approximate end-to-end latency of 20 ms, ensuring that obstacle information remains consistent with the vehicle's current state during real-time operation. The 20 ms figure represents the median end-to-end latency measured during field trials; the observed 95th-percentile latency was 34 ms, with occasional spikes up to 52 ms due to transient congestion in the ROS 2 message queue. To account for this variability, a worst-case latency margin of 50 ms was incorporated into the MPC constraint update logic: occupancy-grid cells flagged as occupied are retained for at least one additional prediction step (i.e., 100 ms) before being cleared. This conservative retention strategy ensures that brief latency spikes do not cause premature relaxation of constraints, thereby maintaining safety margins even under degraded communication conditions. Future work will investigate deterministic scheduling via ROS 2 real-time executors (rclcpp::executors::StaticSingleThreadedExecutor) to further reduce worst-case jitter. Obstacle constraints are integrated into the MPC framework by augmenting the standard optimization problem with inequality constraints derived from the occupancy grid. Specifically, grid cells classified as occupied are translated into spatial exclusion zones within the prediction horizon. At each receding-horizon step, the MPC solver dynamically updates these constraints, allowing the controller to plan and execute smooth avoidance maneuvers without violating vehicle dynamics or actuator limits. Experiments conducted with both static and dynamic obstacles — including moving cones simulating pedestrian behavior — demonstrated the effectiveness of this integration. The MPC controller successfully detected and avoided all dynamic obstacles in real time, with the transient increase in lateral deviation during avoidance maneuvers limited to an average of 0.48 m. This result confirms that the LiDAR-based constraint integration preserves trajectory tracking accuracy while ensuring safe navigation in dynamic environments.

VI. IMPLEMENTATION AND EXPERIMENTAL STUDY

An MPC-based path-tracking controller was implemented for an autonomous golf vehicle operating within the ROS 2 Humble Hawksbill framework. As shown in Figure 1, the system consists of three layers:

- Perception Layer – Maps the environment using 3D LiDAR and constructs an occupancy grid.
- Decision and Control Layer – Computes optimal steering and speed commands using MPC.
- Communication Layer – Provides real-time operation and vehicle interface via Controller Area Network (CAN-BUS).

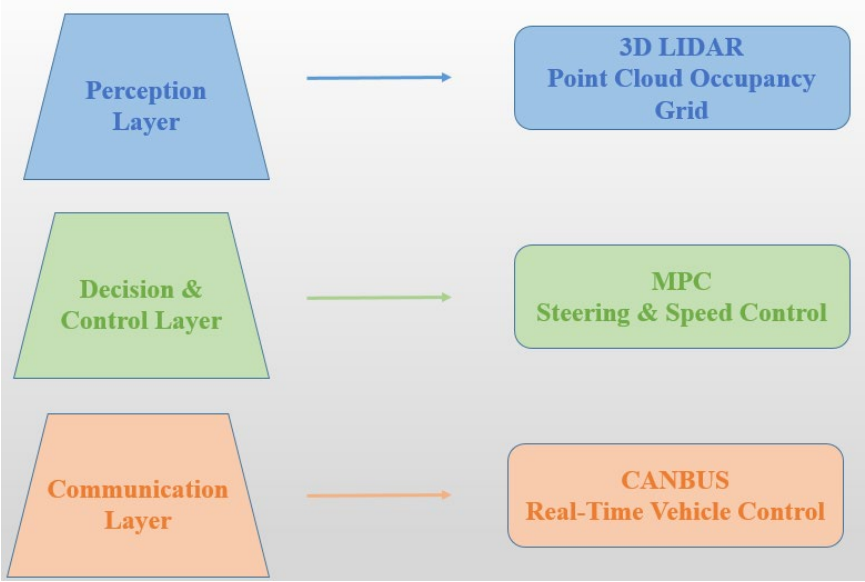


Figure 1. System Layers

The algorithm was implemented on the autonomous golf vehicle shown in Figure 2. The platform is equipped with GNSS, IMU, and 3D LiDAR sensors. The MPC controller runs on a DS-1400 Series rugged embedded computer equipped with a 12th/13th/14th Generation Intel® Core processor (up to 65 W Thermal Design Power (TDP)), 96 GB DDR5 Error-Correcting Code (ECC) memory support, and Peripheral Component Interconnect Express (PCIe) expansion capability. Steering and speed commands computed by the MPC are transmitted to the vehicle via the CAN-BUS interface, and sensor data is fed to the control loop with an approximate latency of 20 ms.

Experiments were conducted on a 350 m closed-loop asphalt path at the MPG Inc. test facility in Konya, Turkey. The route includes three curve sections with radii of approximately 8 m, 12 m, and 20 m, and two straight segments of 60 m and 110 m. The maximum allowable vehicle speed was set to 7 m/s (25.2 km/h). A total of 85,034 position samples were collected over five consecutive laps. Both the MPC and the Stanley controllers were evaluated on the same route under identical weather conditions to ensure comparability. The autonomous vehicle is assigned the mission to travel from point A to point B. Using GNSS coordinates of the defined waypoints, an optimal global path, as shown in Figure 3, is generated and tracked using the MPC algorithm.

Performance evaluation is based on the following metric:

Lateral Error (ey): Defined as the perpendicular distance between the vehicle's actual position and the nearest point on the reference path. This parameter indicates deviation from the centerline.

A total of 85,034 samples were analyzed, and the summary results are presented in Table 1.

Table 1. Lateral Error Performance of the MPC

Metric	**Output**	
	Value	**Description**
Mean Lateral Error (Mean e_y)	0.302	The vehicle maintained stable path tracking within approximately ±30 cm of the reference path throughout the experiment. Std. dev.: 0.381 m.
Maximum Lateral Error (Max e_y)	5.750	Represents a brief transient deviation during entry into a high-curvature turn. The overshoot originated from a sudden maneuver, and the MPC controller recovered the trajectory within approximately 1.2 seconds.
Root Mean Square Error (RMS e_y)	0.447	Reflects system stability; the relatively low RMS value indicates a smooth control response and minimal oscillation. Std. dev.: 0.392 m.

The heading error (e_θ) is defined as the difference between the vehicle's heading and the direction of the reference path. This metric indicates whether the vehicle maintains correct orientation while following the trajectory. As shown in Figure 3, the "Global Path vs. MPC Path" graph shows that the MPC-generated path almost perfectly overlaps the reference trajectory. Furthermore, the consistency between the odometry data and the MPC trajectory confirms that the heading error remained negligible throughout the experiment, demonstrating that MPC successfully achieved accurate vehicle heading control.

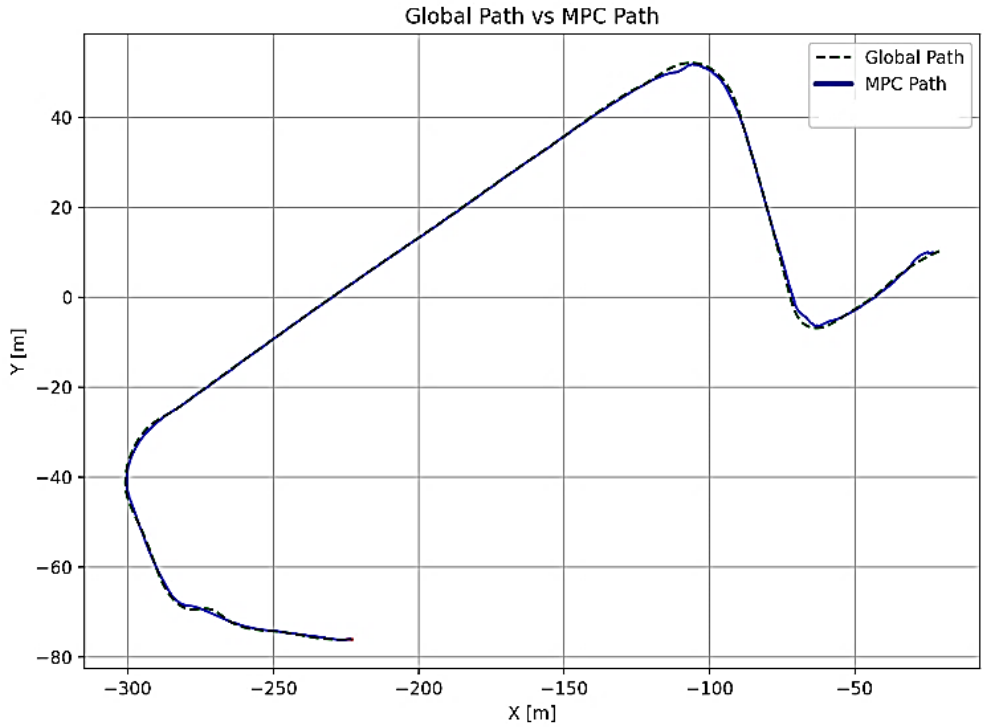


Figure 3. Comparison of Global Path and MPC Path

Experiments were conducted in an open area containing both static and dynamic obstacles. The dynamic obstacles, consisting of moving cones that mimicked pedestrians, were detected in real-time. MPC controller instantly incorporated the updated constraints from the LiDAR-based occupancy grid into the optimization problem, successfully executing safe avoidance maneuvers. The transient increase in lateral deviation during obstacle avoidance maneuvers was limited to an average of 0.48 m. Results show that the MPC controller maintained a mean lateral deviation of ~0.3 m and exhibited negligible heading errors, achieving accurate, stable, and safe trajectory tracking. To ensure a fair comparison, the Stanley controller gain k was tuned via a grid search over $k \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ during a separate calibration run on the same route. The value $k = 1.5$, which yielded the lowest mean lateral error on the calibration run, was selected for the evaluation experiments reported here. The softening parameter $k_s = 1.0$ m/s was kept at its standard value. Heading gain was not tuned separately, as the standard Stanley formulation couples lateral and heading corrections through a single gain. This procedure ensures that the Stanley controller operates at its best achievable performance for this specific route, making the comparison with the MPC representative rather than trivially favorable. For comparison purposes, a Stanley controller was also tested on the same route. While Stanley produced a mean lateral error of 0.51 m and an RMS error of 0.72 m, MPC reduced these values to 0.302 m and 0.447 m, respectively, representing improvements of approximately 41% and 38%. This comparison is summarized in Table 2.

VII. CONCLUSIONS

This study highlights the effectiveness of MPC for path tracking in autonomous vehicles. MPC provides predictive control, explicit constraint handling, and integrated longitudinal/lateral management. Combined with sensor fusion, it enables robust obstacle avoidance and accurate trajectory tracking. In terms of control signals,

the MPC kept the steering angle within ±30° and produced a smooth, oscillation-free control signal; speed commands varied gradually within the 0–7 m/s range.

Challenges remain in computational cost, nonlinear modeling, and real-time optimization, which suggests that future research should focus on:

- Faster and more efficient solvers for real-time applications.
- Integration with machine learning and artificial intelligence to improve robustness under uncertainty.

In summary, MPC offers a powerful and flexible framework for autonomous vehicles, and ongoing interdisciplinary advances are expected to enhance its performance in real-world scenarios further.

Table 2. Controller Performance Comparison

Metric	Controller		
	Value	Stanley	Improvement
Mean Lateral Error (m)	0.302	0.510	~41%
RMS Lateral Error (m)	0.447	0.720	~38%
Max Lateral Error (m)	0.381	-	-

ACKNOWLEDGEMENT

The authors would also like to thank the MPG Machinery Production Group Inc. Co. and the Konya Technical University RAC-LAB Research Laboratory (www.rac-lab.com).

REFERENCES

[1]. R. Yu, H. Guo, Z. Sun, and H. Chen, “MPC-based regional path tracking controller design for autonomous ground vehicles,” in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 2510–2515.
[2]. C. Shen, H. Guo, F. Liu, and H. Chen, “MPC-based path tracking controller design for autonomous ground vehicles,” in *2017 36th Chinese Control Conference (CCC)*, 2017, pp. 9584–9589.
[3]. G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Aggressive driving with model predictive path integral control," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2016, pp. 1433–1440.
[4]. H. Zhou, L. Güvenç, and Z. Liu, “Design and evaluation of path following controller based on MPC for autonomous vehicle,” in *2017 36th Chinese Control Conference (CCC)*, 2017, pp. 9934–9939.
[5]. P. Stano, U. Montanaro, D. Tavernini, M. Tufo, G. Fiengo, L. Novella, and A. Sorniotti, “Model predictive path tracking control for automated road vehicles: A review,” *Annual Reviews in Control*, vol. 55, pp. 194–236, 2023.
[6]. Q. Yao and Y. Tian, “A model predictive controller with longitudinal speed compensation for autonomous vehicle path tracking,” *Applied Sciences*, vol. 9, no. 22, p. 4739, 2019.
[7]. R. C. Colter, “Implementation of the Pure Pursuit Path Tracking Algorithm,” 2007, p. 13.
[8]. D. H. Lee, “Lane-Keeping Control of Autonomous Vehicles Through a Soft-Constrained Iterative LQR,” *IEEE Transactions on Vehicular Technology*, vol. 74, no. 4, pp. 5610–5623, 2025.

[9]. T. Mercy, R. Van Parys, and G. Pipeleers, "Spline-based motion planning for autonomous guided vehicles in a dynamic environment," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 6, pp. 2182–2189, 2017.

[10]. S. Yu, M. Hirche, Y. Huang, H. Chen, and F. Allgöwer, "Model predictive control for autonomous ground vehicles: a review," *Autonomous Intelligent Systems*, vol. 1, no. 1, p. 4, 2021.

[11]. G. Park and S. B. Choi, "A model predictive control for path tracking of electronic four-wheel drive vehicles," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 11, pp. 11352–11364, 2021.

[12]. B. Ahmadi, M. W. Mehrez, W. W. Melek, and A. Khajepour, "Model Predictive Control for Reliable Path Following with Application to the Autonomous Vehicle and Considering Different Vehicle Models," in *2021 5th International Conference on Vision, Image and Signal Processing (ICVISP)*, 2021, pp. 27–32.

[13]. W. Mao, C. Zhang, and H. Zhang, "Design of Path Tracking Control for Wheeled Mobile Robots Based on Model Predictive Control," in *2022 5th International Conference on Intelligent Autonomous Systems (ICoIAS)*, 2022, pp. 238–243.

[14]. Y. Lu, J. Zhu, Z. Wang, N. Du, and J. Zhang, "Path Preview Tracking for Autonomous Vehicles Based on Model Predictive Control," in *2022 IEEE International Conference on Unmanned Systems (ICUS)*, 2022, pp. 405–410.

[15]. M. Wang, J. Chen, H. Yang, X. Wu, and L. Ye, "Path tracking method based on model predictive control and genetic algorithm for autonomous vehicle," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 4661401, 2022.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

From Judgment to Benchmark: An AI-Assisted Wedge Classification Framework for Commercial Loan Spread Governance

Anandasubramanian Pranatharthy Codangudi
PhD Candidate
VELS University
Chennai, India.
anandcp@mac.com +1-470-908-3680
ORCID: 0009-0009-5980-7101

Abstract
When a bank builds a model to estimate the spread a commercial loan should carry, every loan has two prices: the one the lender charged and the one the model expected. The gap between them is not automatically a mistake, and prior work supplies no rule for what to do with it. This paper defines that gap as the pricing wedge. It gives it one: a wedge matters only if it is both statistically unusual and economically large, and it is acceptable only if a documented, evidenced reason explains it. Each loan then routes to one of four outcomes, and a portfolio layer watches for patterns no single loan can show. On 6,000 synthetic loans with planted deviations, this routing escalated 7.6 percent of the book at 0.912 precision, where a simple threshold escalated 17.8 percent at 0.606; the portfolio layer caught planted drift in every run. All evidence is synthetic. Nothing here is demonstrated on real loans.

Keywords: commercial loan pricing, pricing wedge, model risk governance, conformal prediction, explainable AI, fair lending, exception management

I. INTRODUCTION

A commercial lender pricing a corporate facility weighs expected credit loss against funding cost, capital usage, liquidity, collateral, tenor, covenants, fee income, competition, relationship value, and balance-sheet strategy. The resulting spread is a negotiated judgment, not a derivation. When a bank introduces a model that estimates a benchmark spread from observable features, it creates a comparison that did not previously exist: the difference between what the lender charged and what the model would have charged. That difference is the institutional problem. It is not automatically an error, a violation, or evidence of bias. It may reflect documented relationship economics, a verified competitive match, fee-spread substitution, collateral or covenant structure outside the feature set, or a limitation of the model itself. It may equally reflect undocumented concession, pricing drift, or discretion that concentrates in ways a fair-lending examiner would question. Treating every deviation as an exception overwhelms review capacity; treating none as an exception forfeits the control the benchmark was built to provide. The research gap is specific. The literature supplies each component of a response separately. Relationship-banking research establishes why discretion is economically valuable [1–4]. Risk-based pricing research specifies what a defensible benchmark should contain [5], [6]. Machine-learning credit research establishes benchmark feasibility and its distributional risks [7], [8]. Explainability research produces local attributions and documents their manipulability [9-12]. Conformal prediction supplies distribution-free bands [13–16]. Supervisory guidance specifies model governance [17], [18]. None of these treats the model-human deviation as an object with its own definition, classification rule, routing logic, and monitoring layer. The research question follows: can the deviation between a model-based benchmark spread and an executed commercial loan spread be formalized and routed such that justified discretion is preserved, unexplained deviation becomes visible, and review capacity is spent proportionally? This paper makes one contribution. The individual components—gradient boosting, quantile regression, conformal calibration, local attribution, reason-code taxonomies, portfolio monitoring—are

established techniques and are not claimed as novel. The contribution is their formal integration into a single governance framework centered on the pricing wedge, together with a reproducible synthetic study that measures what each architectural layer adds and what it costs. What is validated and what is not must be stated at the outset. We validate the framework's internal behavior on synthetic data with planted ground truth: whether the routing logic separates planted justified from planted unjustified deviations, whether the portfolio layer detects planted drift, and how sensitive both are to thresholds, bandwidth, and reason-code gaming. We do not validate the framework on real commercial loan portfolios, do not establish cross-institutional or cross-jurisdictional generalizability, and do not claim that any measured quantity establishes fair-lending compliance. Section II positions the framework against prior work. Section III defines the framework formally. Section IV specifies the synthetic validation design. Section V reports results. Section VI presents the governance protocol, reason-code taxonomy, and fairness-control specification; Section VII addresses generalizability and limitations. Section VIII concludes.

II. RELATED WORK AND CONTRIBUTION GAP

1. Commercial Loan Pricing and Relationship Banking

Petersen and Rajan [1] and Berger and Udell [2] established that lending relationships generate private information that improves credit availability and contract terms. Both study small-business lending, and the transfer of their findings to middle-market and large corporate lending requires justification rather than assumption. That justification exists in the corporate-loan literature: Bharath, Dahiya, Saunders, and Srinivasan [3] find, on syndicated loans to public borrowers, that repeated borrowing from the same lender lowers spreads by roughly 10 to 17 basis points, and that relationship value is concentrated where borrower transparency is low, becoming statistically indistinguishable for the largest borrowers. Related work documents what the lender captures from the relationship [4]. This bounds the present framework's scope: relationship-based deviation is an expected feature of opaque middle-market credit and a weaker expectation for large, transparent borrowers, so a single institution-wide materiality threshold is unlikely to be appropriate across segments.

2. Risk-Based Pricing and Bank Economics

Loan spreads reflect bank-side factors as well as borrower risk. Berg, Saunders, and Steffen [5] show that most U.S. syndicated loans carry fee structures alongside spreads, and that fees price embedded options and screen borrowers. The direct implication for wedge measurement is that a spread deviation observed without the fee structure can be economically misleading: fee-spread substitution can produce a large negative wedge on an economically unchanged transaction. Bank organizational structure also affects the capacity to act on soft information [6], which is why a centrally calibrated benchmark may systematically diverge from locally optimal pricing.

3. Machine Learning in Credit and Pricing

Machine-learning models improve predictive accuracy in credit applications, and their distributional consequences are documented. Fuster, Goldsmith-Pinkham, Ramadorai, and Walther [7] show, using U.S. mortgage data, that flexible models change the dispersion of predicted risk across borrower groups, with some groups disproportionately unlikely to gain. This establishes that a benchmark's accuracy and its distributional properties are separate questions, and that improving the first does not settle the second.

4. Explainability and Its Limitations

SHAP [9] and LIME [10] produce local attributions that make a single benchmark estimate reviewable. Their limitations are equally established. Slack, Hilgard, Jia, Singh, and Lakkaraju [11] demonstrate that perturbation-based post-hoc explanations can be deliberately manipulated so that a biased model yields innocuous explanations, and Kumar et al. [12] show that Shapley-value attributions do not straightforwardly answer the questions practitioners ask of them. Rudin [19] argues that for high-stakes decisions, an interpretable model is preferable to a post hoc explanation of an opaque model. This paper adopts the implication rather than disputing it: explanation is treated as an input to human review under a documented evidence standard, never as proof that a benchmark is correct, and the model comparison in Section V tests whether an opaque model is needed at all.

5. Uncertainty Quantification and Conformal Prediction

Conformal prediction produces prediction intervals with finite-sample marginal coverage without distributional assumptions [13]. Conformalized quantile regression [14] adapts the interval width to heteroscedasticity, which matters here because pricing dispersion is not constant across borrower segments. Two limitations are load-bearing for governance. Marginal coverage does not imply conditional coverage: a method can meet its nominal level overall while under-covering a specific segment, which motivates the equalized-coverage diagnostics of Romano, Barber, Sabatti, and Candès [15]. And coverage guarantees rest on exchangeability, which fails under distribution shift [16], [20]. A pricing book subject to competitive and macroeconomic movement is exactly such a setting, so band validity is a monitored property, not a standing assumption.

6. Algorithmic Discrimination and Fair-Lending Monitoring

Bartlett, Morse, Stanton, and Wallace [8] estimate, on U.S. mortgage data, that risk-equivalent minority borrowers paid measurably higher rates, and that algorithmic lenders exhibited disparities that were similar for some products and lower for others. The relevant result here is not the point estimate but the mechanism: pricing disparities can persist in data subsequently used to train a model. A benchmark fitted to that history encodes it. The framework, therefore, treats the benchmark as a potential carrier of inherited bias. It measures that inheritance directly (Section V), rather than presenting the benchmark as a neutral standard against which human pricing is judged.

7. Model-Risk Governance

U.S. interagency supervisory guidance on model risk management was revised on 17 April 2026 by SR 26-2 [17], which supersedes SR 11-7 (2011) [21] and the 2021 interagency statement, and is accompanied by FDIC FIL-15-2026 [18] and the corresponding OCC bulletin. The revised guidance retains the core expectations of development, independent validation, and ongoing monitoring, and articulates them on a risk-based, proportional basis tailored to an institution's model risk profile, size, and complexity; it is expected to be most relevant to banking organizations with more than $30 billion in total assets. Supervisory guidance does not have the force and effect of law and does not create binding legal obligations; it describes supervisory expectations. This paper cites SR 26-2 as the operative reference and SR 11-7 only for historical continuity.

8. Contribution Relative to Prior Approaches

Table 1 compares the framework against the governance approaches a commercial bank already operates. The comparison is across four dimensions: which object each approach governs, whether it uses an uncertainty-

aware materiality test, whether it imposes a structured evidence standard on human justification, and whether it aggregates individually non-material outcomes at the portfolio level. No existing approach is deficient at its own task; the gap is that none takes the model-human deviation as its object.

Table 1. Comparison with Existing Governance Approaches

Approach	**Governed object**	**Uncertainty-aware materiality**	**Structured evidence standard**	**Portfolio aggregation of non-material items**
Pricing-exception policy	Deviation from the internal pricing rule	No: fixed threshold	Partial: free-text justification common	No
Fair-lending pricing monitoring	Outcome disparity across groups	No	No	Yes, but not linked to model deviation
Model-risk governance [17]	Model design, use, and performance	Indirect, via validation	No: governs the model, not the deviation	No
Explainable-AI workflow [9], [10]	Model output attribution	No	No	No
Human-in-the-loop review	Algorithmic recommendation	No	Varies by institution	No
Proposed wedge framework	**Model–human pricing deviation**	**Yes: band and threshold jointly**	**Yes: four required elements**	**Yes: drift and segment layer**

III. METHODOLOGY: THE PRICING-WEDGE FRAMEWORK

1. Formal Problem Definition

Let loan i be priced at time t. The institution observes a feature vector $x_{i,t}$ that contains borrower risk, facility structure, bank cost, and market variables available at or before the pricing decision. The lender executes an actual spread; a model estimates a benchmark spread and an interval around it. The governance problem is to map each loan to exactly one of four outcomes—retain record, monitor at portfolio level, document and approve, escalate for review—using only quantities that can be computed, documented, and audited.

2. Actual Spread

The actual spread $s^{a}_{i,t}$ is the executed contractual margin over the relevant reference rate, in basis points. Where material fees exist, the spread must be interpreted together with the fee structure, because fee-spread substitution can make an economically unchanged transaction appear as a large negative deviation [5].

3. Benchmark Spread

The benchmark spread $\hat{s}^{b}_{i,t} = f(x_{i,t})$ is a supervised estimate of the spread the institution's own pricing function would produce for comparable observable characteristics. It is a model output subject to model-risk

governance [17]. It is not objective, optimal, or sovereign, and it does not observe unrecorded soft information, undocumented relationship value, or competitive intelligence outside the feature set.

4. Pricing Wedge

The pricing wedge is defined as:

$$w_{i,t} = s^{a}_{i,t} - \hat{s}^{b}_{i,t} \quad (1)$$

expressed in basis points, with sign preserved. A positive wedge means the loan was priced above the benchmark; a negative wedge means it was priced below the benchmark. Sign is retained throughout because the two directions raise different governance questions: materially negative wedges raise underpricing and undocumented-concession questions, materially positive wedges raise overpricing and fairness questions.

5. Expected Pricing Band

Let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ be quantile regression estimates fitted on a training partition, and let E be the conformity score computed on a disjoint calibration partition of size n, with $Q_{1-\alpha}(E)$ its $\lceil (n+1)(1-\alpha) \rceil$-th order statistic [14]. The expected pricing band is:

$$[\hat{L}_{i,t}, \hat{U}_{i,t}] = [\hat{q}_{\alpha/2}(x_{i,t}) - Q_{1-\alpha}(E), \hat{q}_{1-\alpha/2}(x_{i,t}) + Q_{1-\alpha}(E)] \quad (2)$$

This construction delivers marginal coverage at level $1-\alpha$ under exchangeability. It does not deliver conditional coverage for an arbitrary segment, and its guarantee degrades under distribution shift [15], [16]. Both properties are monitored quantities in Section VI rather than assumptions.

6. Hybrid Materiality Rule

A wedge is material only when it is both statistically unusual and economically significant. With θ the institution-approved business threshold in basis points:

$$M_{i,t} = \mathbf{1}[\, s^{a}_{i,t} \notin [\hat{L}_{i,t}, \hat{U}_{i,t}] \,] \cdot \mathbf{1}[\, |w_{i,t}| > \theta \,] \quad (3)$$

The first condition prevents small deviations from being escalated where the model is confident, but the deviation is economically trivial. The second prevents large deviations from being escalated where the model itself is highly uncertain. The conjunction is binding only when θ is large relative to the bandwidth; Section V measures the range over which it binds.

7. Explainability Rule

Explainability is an operational indicator, not a property of the model. A wedge is explainable when a submitted reason code meets the minimum standard defined in Section VI, which requires all four: a category from the institution-approved taxonomy, a quantified pricing or economic rationale, supporting evidence, and a named decision owner with approval level.

$$X_{i,t} = \mathbf{1}[\text{ code submitted }] \cdot \mathbf{1}[\text{ all four required elements present and evidenced }] \quad (4)$$

A wedge with no code, or with a code that fails the evidence requirement, is unexplained regardless of the decision owner's subjective belief that the deviation was justified.

8. Reason-Code Taxonomy

The taxonomy is a structured input, not a free-text field, because free-text justification cannot be aggregated across a portfolio and therefore cannot support drift or segment monitoring. Table 5 specifies the categories, evidence requirements, approval levels, and the known gaming risk of each.

9. Four-Cell Governance Matrix

Materiality and explainability jointly determine the routing outcome, as set out in Table 2:

$$R(M, X) = \{\text{retain, monitor, approve, escalate}\} \text{ for } (M, X) = (0,1), (0,0), (1,1), (1,0) \quad (5)$$

Table 2. Pricing Wedge Classification Matrix

	Explainable ($X = 1$)	**Unexplained ($X = 0$)**
Non-material ($M = 0$)	Retain record; no loan-level action	Monitor at portfolio level
***M*aterial (M = 1)**	Document and approve at the authority level	Escalate for review

Figure 1 shows the complete routing path from executed price to governance outcome.

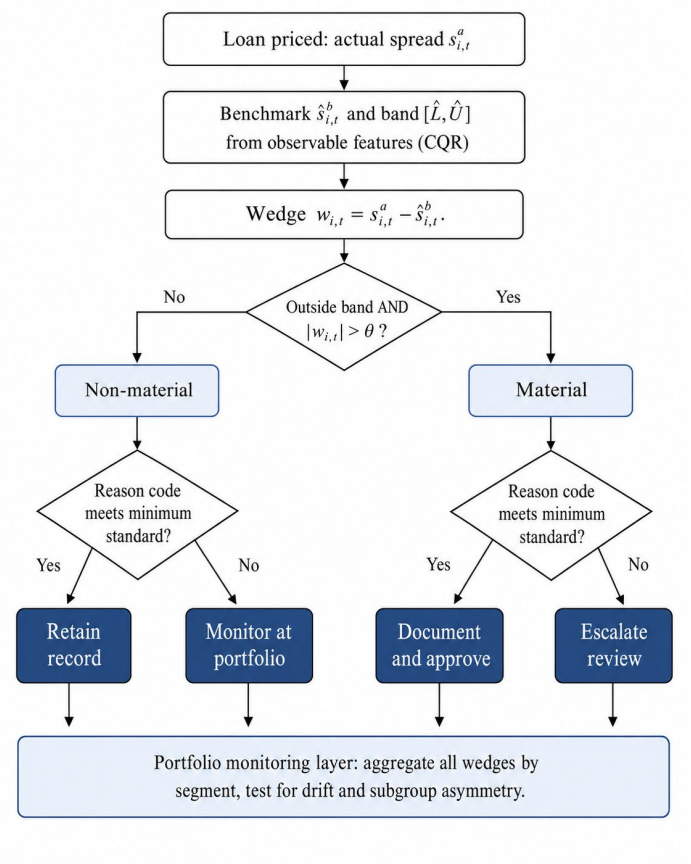


Figure 1. Wedge classification and routing flow.

10. Portfolio-Monitoring Layer

Loan-level routing is necessary but not sufficient. Individually non-material wedges can aggregate into a pattern that no single loan would disclose. For segment *k,* the layer compares the mean wedge in a recent window against a baseline window over loans *not* escalated individually:

$$z_k = (\bar{w}_{k,\,recent} - \bar{w}_{k,\,base}) / (\sigma_{k,\,base} / \sqrt{n_{k,\,recent}}) \quad (6)$$

Segments with |zk| > 2 and a minimum sample size are referred to the governance committee. The statistic is a monitoring trigger for human review, not a hypothesis test with a controlled error rate; Section V reports its false-flag rate, and Section VI specifies the multiple-comparison treatment.

11. Algorithmic Workflow

Algorithm 1 states the complete procedure. It is written so that an institution with a loan-level pricing dataset, a reason-code field, and a segment identifier can implement it without further specification.

Algorithm 1: Wedge classification and routing
Input: loans $D = \{(x_{i,t}, s^a_{i,t}, \text{code}_{i,t}, \text{segment}_{i,t})\}$; level $1-\alpha$; threshold θ
Output: routing outcome per loan; segment drift flags
1. Partition *D* by time into train, calibration, validation, and test.
2. Fit candidate benchmark models on train; select on validation MAE.
3. Fit lower and upper quantile models at $\alpha/2$ and $1-\alpha/2$ on train.
4. Compute conformity scores on calibration; set $Q_{1-\alpha}(E)$ per (2).
5. for each loan *i* in the monitoring population do
6. $\hat{s}^b_{i,t} \leftarrow$ selected model; band $\leftarrow$ (2); $w_{i,t} \leftarrow$ (1)
7. $M_{i,t} \leftarrow$ (3); $X_{i,t} \leftarrow$ (4)
8. $\text{outcome}_{i,t} \leftarrow R(M_{i,t}, X_{i,t})$ per (5)
9. record $\hat{s}^b$, band, *w*, *M*, *X*, outcome, local attribution, owner
10. end for
11. For each segment *k,* over non-escalated loans, do
12. compute z_k per (6); flag if $|z_k| > 2$ and $n_{k,\text{recent}} \geq n_{\min}$
13. end for
14. Review band coverage, segment coverage, threshold, and taxonomy use in the committee.

IV. Synthetic Validation Design

This section specifies a reproducible synthetic study. Its purpose is to measure the framework's internal behavior under conditions where the correct answer is known by construction. Three categories must be kept distinct throughout. Synthetic results describe how the framework behaves on data we generated. Actual commercial-loan evidence is not available to this study and is not presented. Real-world generalizability is not established by anything reported here.

1. Data-Generating Process

We generate 6,000 synthetic commercial loans under fixed seed 20260719. Each loan carries a borrower risk grade on a ten-point scale, log facility size, tenor, secured indicator, covenant strength, sector, region, quarter of origination over twelve quarters, a funding-cost index, competitive intensity, relationship depth, fee-income

ratio, and capital usage. A risk-based spread is formed as a nonlinear function of these variables, including a quadratic risk term and a risk-tenor interaction, plus heteroscedastic noise that defines two uncertainty regimes: a low-dispersion regime with a 13-basis-point standard deviation and a high-dispersion regime with 34 basis points, covering weaker grades and the technology and services sectors.

These ranges are *plausible synthetic design assumptions*. They are not empirically calibrated to any real loan portfolio, and no claim of realism beyond internal plausibility is made for them.

2. Planted Ground Truth

Deviations are then planted with a known type. Justified deviations comprise documented relationship concessions, verified competitive matches, and fee-spread substitution, each carrying an adequate reason code. Unjustified deviations of both signs are planted without an adequate code; a subset of these submits a relationship-value code that fails the evidence requirement, providing the reason-code gaming scenario. A segment-level markup is applied to a monitoring segment defined by region and facility size, so that the segment correlates with observable features, and a benchmark trained on this history can absorb part of the markup. Finally, small individually non-material concessions are planted in one sector in the final two quarters, providing the portfolio-drift scenario. Ground truth for loan-level evaluation is defined as an unjustified deviation in which the planted magnitude exceeds the business threshold.

3. Partitions and Models

The data are split 50 percent train, 20 percent calibration, 15 percent validation, and 15 percent test. Quantile models and benchmark models are fitted on train only; the conformity quantile is computed on calibration only; model selection uses validation only; all reported metrics use the held-out test partition, except the portfolio layer, which runs on all 2,790 non-training loans because portfolio monitoring operates on the whole book rather than a sample.

Two benchmark candidates are compared for a stated reason. A transparent linear model represents the interpretable option that Rudin [19] argues should be preferred, where it suffices. A histogram gradient-boosting model [22] represents the flexible option commonly adopted for tabular credit data. The comparison exists to test whether the framework requires an opaque model, not to advocate one. Bands are built using conformalized quantile regression [14] at a 90 percent nominal level, with gradient boosting and the pinball loss [23].

Crucially, benchmark models are trained on *actual* historical spreads that contain planted deviations and planted segment markup. This is deliberate: it reproduces the bias-inheritance condition that the framework must contend with in practice. To measure inheritance, we also fit a counterfactual benchmark on an otherwise identical history in which the segment markup never occurred.

4. Governance Architectures Compared

Four routing architectures are evaluated. Architecture A escalates based solely on the business threshold. Architecture B escalates on the prediction band alone. Architecture C applies the hybrid two-condition rule at the loan level. Architecture D is the full framework: architecture C plus the portfolio-monitoring layer. All four escalate only where the reason code fails the evidence standard. The comparison isolates what each layer adds.

V. RESULTS

The executed analysis produces all figures in this section. Headline values are from the fixed-seed run; dispersion is from twenty independent seeds.

1. Benchmark Accuracy and Model Choice

On validation, the transparent linear model achieved a mean absolute error of 37.57 basis points, compared to 40.56 for gradient boosting, and was selected in all 20 seeds. On the held-out test partition, the linear model recorded 39.82 basis points MAE and 56.88 RMSE, against 41.91 and 59.12 for gradient boosting; across seeds, the gap was 38.71 ± 1.58 versus 41.24 ± 1.47.

We do not claim gradient boosting is preferable here; on this data-generating process, it is not. Pricing noise dominates the nonlinearity we planted, and the flexible model pays a variance cost for curvature it cannot reliably recover. The governance implication is the useful one: the framework does not require an opaque benchmark, and where a transparent model matches a flexible one, the interpretable option should be preferred on review grounds alone [19]. Section V confirms that routing behavior is essentially unchanged when the gradient-boosting benchmark is substituted, which is evidence that the framework is model-agnostic rather than evidence that either model is correct.

2. Uncertainty Bands

At a 90 percent nominal level, the conformalized band achieved 89.0 percent marginal coverage on test and 89.9 ± 1.2 percent across seeds, with a mean width of 192.73 basis points. Conditional coverage was not uniform: 90.5 percent in the low-dispersion regime versus 88.0 percent in the high-dispersion regime, and 89.5 percent versus 87.9 percent across the monitoring segment. Bandwidth did adapt to dispersion, from 184.79 to 198.00 basis points. This is the marginal-versus-conditional distinction in operational form: nominal coverage was met overall while a segment was under-covered by roughly one to two percentage points, which is precisely why subgroup coverage is a standing monitoring item [15].

3. Governance Routing

Table 2 reports the four architectures against planted ground truth. The result is a trade-off, not a dominance.

Table 3. Governance Routing Against Planted Ground Truth (Test Partition)

Architecture	**Precision**	**Recall**	**F1**	**False-escalation rate**	**Share escalated**
A: business threshold only	0.606	0.907	0.727	0.079	17.8%
B: prediction band only	0.912	0.579	0.709	0.008	7.6%
C: hybrid, loan level only	0.912	0.579	0.709	0.008	7.6%
D: full framework (C + portfolio layer)	0.912	0.579	0.709	0.008	7.6%

Threshold-only routing recovered the most planted cases: recall 0.907 versus 0.579 for the hybrid rule, and, across seeds, its F1 was higher (0.712 ± 0.025 versus 0.661 ± 0.051). It achieved this by escalating 17.8 percent of the book at 0.606 precision, so roughly two in five escalations were loans with no planted problem. The hybrid rule escalated 7.6 percent, with 0.912 precision and a false-escalation rate of 0.008, compared to 0.079. Which architecture is preferable is a governance choice about the relative cost of a missed deviation and a wasted review, not a fact that the data settles. Institutions with scarce senior-review capacity should read this

table as favoring the hybrid rule; institutions prioritizing recall should not adopt it without raising review capacity accordingly.

Architectures B and C were identical at this threshold. The two-condition rule is non-binding whenever the band is wide relative to θ: at a mean width of 192.73 basis points, a loan outside the band almost always exceeds the 50-basis-point threshold. Table 4 shows the threshold at which the conjunction begins to bind.

Table 4. Sensitivity to Business Threshold and Band Coverage

Setting	**Mean band width (bp)**	**Precision**	**Recall**	**F1**	**Share escalated**
θ = 25 bp	—	0.912	0.579	0.709	7.6%
θ = 50 bp (baseline)	—	0.912	0.579	0.709	7.6%
θ = 75 bp	—	0.909	0.561	0.694	7.3%
θ = 100 bp	—	0.946	0.495	0.650	6.2%
95% band	248.05	0.950	0.355	0.517	4.4%
90% band (baseline)	192.73	0.912	0.579	0.709	7.6%
80% band	127.20	0.733	0.794	0.762	12.9%

Both calibration choices are governance instruments rather than statistical details. Widening the band from 80 to 95 percent coverage increased the mean width from 127.20 to 248.05 basis points, reduced recall from 0.794 to 0.355, and raised precision from 0.733 to 0.950. A wide band produces a quiet exception queue by construction, which is why bandwidth is itself a reviewable parameter.

4. Portfolio Layer and Drift Detection

The portfolio layer ran on 2,790 non-escalated loans. It detected the planted sector drift in the fixed-seed run at -2.18 standard errors, a mean shift of -9.41 basis points, and detected it in twenty of twenty seeds. Loan-level routing detected none of these loans, and could not: every drift loan was individually below the materiality threshold and carried an adequate code. This is the only dimension on which architecture D differs from architecture C, and it is the specific justification for the portfolio layer.

The layer is not free. In the fixed-seed run, it also flagged a second sector where nothing was planted at 2.56 standard errors, and across seeds it produced 0.25 false segment flags per run (standard deviation 0.43). With five segments tested per run and no multiplicity correction, this is the expected behavior of an uncorrected screen, and this is why Section VI treats segment flags as review triggers requiring investigation rather than as findings.

Figure 2 summarizes loan-level routing and the segment statistics side by side.

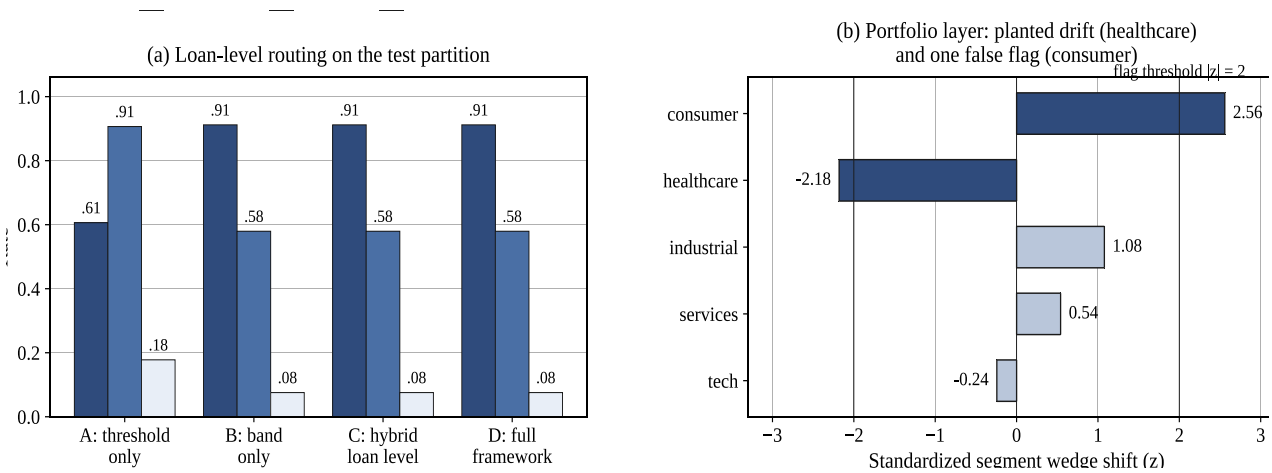


Figure 2. Governance routing performance and portfolio-layer segment statistics.

5. Reason-Code Gaming and Model Substitution

When claimed reason codes were accepted without the evidence requirement, hybrid recall fell from 0.579 to 0.364 and F1 from 0.709 to 0.513, while the escalated share fell from 7.6 to 5.0 percent. The queue became quieter while detection deteriorated. The evidence requirement in the reason-code standard is therefore load-bearing rather than administrative: without it, the framework can be silenced by populating a taxonomy field.

Substituting the gradient-boosting benchmark for the selected linear benchmark left hybrid routing unchanged at 0.912 precision and 0.579 recall.

6. Bias Inheritance

The planted segment markup produced a true wedge gap of 10.52 basis points between the monitoring segment and the rest of the portfolio. The counterfactual benchmark, trained on a history without the markup, surfaced 10.29 basis points of it. The deployed benchmark, trained on the biased history, yielded only 9.09 basis points: it absorbed 1.19 basis points, roughly 12 percent of the detectable disparity, into its predictions. Across seeds, the absorbed amount was 1.32 ± 0.20 basis points.

The magnitude is modest here because the segment was only partly predictable from the feature set; with a segment more strongly correlated with observable features, absorption would be larger. The direction is what matters, and it is unfavorable: a benchmark fitted to biased history understates the disparity that the wedge is supposed to reveal, and it does so silently. The wedge cannot detect bias that the benchmark has already learned. This is a structural limit of the design, not a tuning problem, which is why the fairness controls in Section VI include a review of the training distribution rather than only the wedge distribution.

7. Interpretation and Scope

Three readings must be kept separate. Effects that follow from the design include the detection of planted drift and deterioration under reason-code gaming; these test whether the mechanism does what it is specified to do and would be alarming only if they failed. Findings that demonstrate internal behavior include only the precision and recall levels, the threshold at which the two-condition rule binds, and the false-flag rate; their magnitudes depend on the planted deviation rates and dispersions and would change under different assumptions. Findings that would change materially under different distributions include every absolute rate

reported above; none should be read as a performance expectation for a real portfolio. Nothing here has been tested on real loan data, and no number in this section speaks to fair-lending compliance, which is a legal determination that no metric establishes.

VI. GOVERNANCE PROTOCOL AND CONTROL ARCHITECTURE

The framework becomes operational only inside a repeatable control protocol with named owners. Table 5 states the reason-code taxonomy; the fairness and model-risk controls follow; Figure 3 shows how the components nest.

1. Reason-Code Taxonomy

Table 5. Operational Reason-Code Taxonomy

Category	**Required evidence and quantification**	**Approval level**	**Review or expiry**	**Known gaming risk**
Documented relationship economics	Cross-product revenue reconciliation from the relationship system; quantified annual value	Relationship manager plus credit officer	Annual; expires if realized revenue misses projection	Highest: unfalsifiable if projections are never reconciled
Verified competitive match	Dated competing term sheet or documented market quote; quantified spread differential	Credit officer	Per transaction	Unverifiable verbal quotes
Fee-spread substitution	Fee schedule and total-cost calculation showing economic equivalence [5]	Pricing desk	Per transaction	Low if total cost is computed; high if asserted
Collateral or covenant difference not in the feature set	Executed documentation reference and structural description	Credit officer	Until amendment	Recurrence signals a feature-set gap, not an exception
Strategic acquisition	Named strategic mandate; expected relationship trajectory	Senior credit committee	Twelve months, then re-justify	Open-ended use as a residual category
Model limitation	Attribution report showing the driver the benchmark mis-estimates	Model risk plus credit officer	Each model recalibration	Concentration indicates the model, not the loan, requires action
Data-quality issue	Identified field, corrected value, remediation ticket	Data owner	Until corrected	Persistent use masks unresolved data defects
Other	Written rationale plus quantification; enhanced review mandatory	Senior credit committee	Per transaction	Residual bucket absorbing unclassifiable discretion



Two categories are deliberately constrained. Relationship value may not serve as an unbounded explanation: it requires quantification at submission and reconciliation with realized revenue, and unreconciled claims by a single manager constitute a portfolio-monitoring field in their own right. Management judgment is not a permitted category. Section V shows why: accepting claimed codes without evidence cuts detection by roughly a third.

2. Fairness-Control Specification

The unit of analysis is the individual loan, aggregated to monitoring segments. Comparison groups are defined only by variables the institution may lawfully use for monitoring in its jurisdiction; the synthetic study uses a geographic and facility-size segment, and no demographic variable is used or implied. Conditioning variables are the risk, structural, and bank-cost features in the benchmark feature set, so that segment comparisons are made among comparable loans rather than across the raw book. The reviewed quantities are conditional mean wedge differences, material wedge rate differences, escalation rate differences, reason code adequacy rate differences, and subgroup band coverage.

Pricing-exception monitoring is itself a long-standing supervisory focus in fair-lending examination [24], so the segment diagnostics above are an extension of an existing control rather than a new reporting obligation. Review runs quarterly. Segments with fewer than the minimum sample defined by the model-risk policy are pooled or carried forward rather than reported. Because several segments are screened simultaneously, the false-flag behavior measured in Section V applies: flags are treated as triggers for investigation, a false-discovery-rate correction is applied where the number of segments is large, and no flag is recorded as a finding before investigation. Escalation goes to the fair-lending and model-risk committee, which must document the investigation, the conclusion, and any changes to thresholds, models, or policies. Implementation must be adapted to applicable law, data availability, jurisdiction, and institutional policy; no metric in this framework establishes legal compliance.

3. Model-Risk Controls and Decision Ownership

The benchmark operates within the institution's model-risk framework, scaled to its model-risk profile, as SR 26-2 directs [17]. Required controls include a challenger model maintained in parallel; independent validation at deployment and at policy-defined intervals; documented recalibration procedures that assess whether recalibration propagates inherited bias; monitoring of band coverage, both overall and by segment; and a complete audit trail of model versions, thresholds, classifications, reason codes, and outcomes. Local attribution accompanies every benchmark estimate in basis points, so a reviewer can identify which features drove the estimate; consistent with [11] and [12], attribution supports review and never substitutes for validation.

Decision ownership is unchanged by any of this. The lender and credit authority own the pricing and approval decision. The framework changes what must be explained, not who decides. Figure 3 shows the resulting arrangement: the benchmark sits inside the control layers rather than above them.

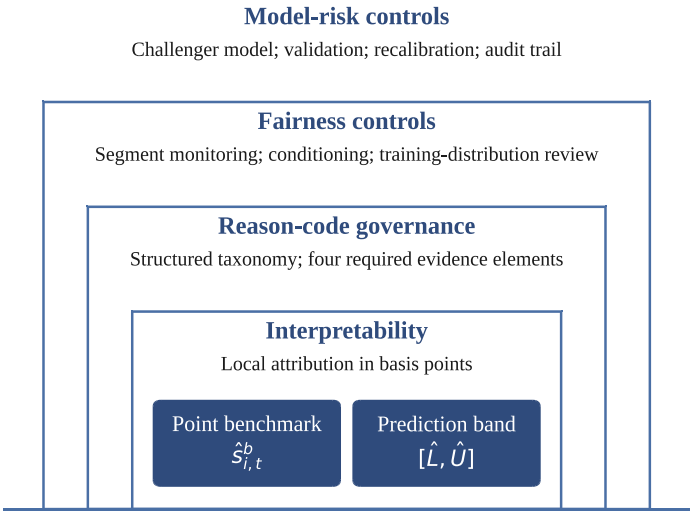


Figure 3. Control architecture around the benchmark.

VII. DISCUSSION, GENERALIZABILITY, AND LIMITATIONS

1. What the Synthetic Study Does and Does Not Establish

The study establishes that the specified logic behaves as specified: the routing rule separates planted justified from planted unjustified deviations at measured precision and recall, the portfolio layer detects aggregate drift that loan-level routing structurally cannot, the evidence requirement is load-bearing, and a benchmark fitted to biased history absorbs part of the disparity it is meant to reveal. It establishes none of the following: that the framework improves governance quality at a real bank, that the measured rates transfer to any real portfolio, or that the planted deviation rates resemble those of any institution. Synthetic ground truth is a test of internal consistency, not of external validity. Real loan-level validation remains necessary as the next step and requires proprietary pricing data, reason-code records, and outcome data that this study lacked.

2. Generalizability Across Banks, Portfolios, and Jurisdictions

Three heterogeneities bound transfer. Bank heterogeneity: institutions differ in pricing-model maturity, review capacity, and data integration, and the architecture choice in Table 2 is capacity-dependent rather than universal. Portfolio heterogeneity: relationship-based deviation is expected in opaque middle-market credit and much weaker for large transparent borrowers [3], so a single institution-wide threshold will over-escalate one segment while under-escalating another; thresholds should be segment-specific. Jurisdictional heterogeneity: permissible monitoring variables, disclosure obligations, and supervisory expectations differ, and the supervisory references here are U.S. Nothing in the mechanism is U.S.-specific, but the fairness-control specification must be rewritten for each jurisdiction.

Applicability also has a lower bound in operational cost. The framework requires loan-level pricing data joined to credit, funding, fee, relationship, and approval systems, plus a structured reason-code field enforced in the pricing workflow. Institutions without that integration cannot implement it as specified, and partial implementation without the portfolio layer forfeits the one capability that distinguishes it from a conventional exception policy.

3. Structural Limitations

Relationship value remains partly unobservable. The framework cannot verify, on a case-by-case basis, that a documented claim reflects genuine economics; it can require that the claim be documented and tested against realized revenue at the portfolio level. That defense operates after loans are booked, not at the moment of decision.

Reason codes can be gamed, and Section V quantifies the cost. Structured codes solve aggregation, not sincerity. The defense is portfolio-level monitoring of code distributions and outlier review of high-frequency submitters.

The benchmark may inherit bias, and Section V shows that it does so measurably. The fairness layer addresses this through segment monitoring, conditioning, and training-distribution review, but no control removes inherited bias; each only limits how far it propagates undetected.

Materiality thresholds and bandwidth are governance choices with measurable consequences, not neutral parameters. Table 4 shows that a threshold set too high leaves material unexplained and wedges unescalated. At the same time, a band set too wide produces a quiet queue that may reflect model weakness rather than pricing discipline. Both require periodic stress-testing against review capacity.

Finally, model misspecification and distribution shift bound the whole construction. If the benchmark omits a driver that the market prices, the wedge measures the omission rather than the discretion; if the pricing distribution shifts, the conformal guarantee degrades [16], [20], and coverage monitoring is the only available warning.

VIII. CONCLUSION

Commercial loan pricing is a judgment based on risk, funding costs, capital, competition, and relationship value. When a bank adds a model-based benchmark, it creates a difference between what was charged and what the model expected, and that difference is not self-interpreting.

This paper defined that difference as the pricing wedge and built a governance framework around it: a conformal band and a business threshold that jointly decide materiality, a reason-code standard that decides explainability, a four-cell rule that routes each loan to one of four outcomes, and a portfolio layer that aggregates what loan-level review cannot see.

On 6,000 synthetic loans with planted ground truth, hybrid routing escalated 7.6 percent of the book at 0.912 precision where threshold-only routing escalated 17.8 percent at 0.606; the portfolio layer detected planted drift in every seed while loan-level routing detected none of it; accepting unevidenced reason codes cut detection by roughly a third; and a benchmark trained on biased history absorbed part of the disparity it was meant to expose. These are properties of the framework measured on data we generated. They are not evidence about any real portfolio.

The contribution is architectural: the integration of benchmark estimation, uncertainty-aware materiality, structured explainability, proportional routing, and portfolio monitoring into one governed object. The principal limitation is that the evidence is synthetic. The next step is validation on real loan-level pricing and reason-code data at an institution willing to measure classification stability and review outcomes over successive governance cycles.

REFERENCES

[1]. M. A. Petersen and R. G. Rajan, "The benefits of lending relationships: Evidence from small business data," Journal of Finance, vol. 49, no. 1, pp. 3–37, 1994.

[2]. A. N. Berger and G. F. Udell, "Relationship lending and lines of credit in small firm finance," Journal of Business, vol. 68, no. 3, pp. 351–381, 1995.
[3]. S. T. Bharath, S. Dahiya, A. Saunders, and A. Srinivasan, "Lending relationships and loan contract terms," Review of Financial Studies, vol. 24, no. 4, pp. 1141–1203, 2011.
[4]. S. T. Bharath, S. Dahiya, A. Saunders, and A. Srinivasan, "So what do I get? The bank's view of lending relationships," Journal of Financial Economics, vol. 85, no. 2, pp. 368–419, 2007.
[5]. T. Berg, A. Saunders, and S. Steffen, "The total cost of corporate borrowing in the loan market: Don't ignore the fees," Journal of Finance, vol. 71, no. 3, pp. 1357–1392, 2016.
[6]. A. N. Berger, N. H. Miller, M. A. Petersen, R. G. Rajan, and J. C. Stein, "Does function follow organizational form? Evidence from the lending practices of large and small banks," Journal of Financial Economics, vol. 76, no. 2, pp. 237–269, 2005.
[7]. A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably unequal? The effects of machine learning on credit markets," Journal of Finance, vol. 77, no. 1, pp. 5–47, 2022.
[8]. R. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-lending discrimination in the FinTech era," Journal of Financial Economics, vol. 143, no. 1, pp. 30–56, 2022.
[9]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.
[10]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[11]. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2020, pp. 180–186.
[12]. I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in Proc. 37th Int. Conf. Machine Learning (ICML), 2020.
[13]. V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY: Springer, 2005.
[14]. Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in Advances in Neural Information Processing Systems 32, 2019, pp. 3538–3548.
[15]. Y. Romano, R. F. Barber, C. Sabatti, and E. J. Candès, "With malice toward none: Assessing uncertainty via equalized coverage," Harvard Data Science Review, vol. 2, no. 2, 2020.
[16]. R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in Advances in Neural Information Processing Systems 32, 2019.
[17]. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, and Federal Deposit Insurance Corporation, "Revised guidance on model risk management," SR Letter 26-2, April 17, 2026.
[18]. Federal Deposit Insurance Corporation, "Agencies revise the interagency model risk management guidance," Financial Institution Letter FIL-15-2026, April 17, 2026.
[19]. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
[20]. R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," Annals of Statistics, vol. 51, no. 2, pp. 816–845, 2023.
[21]. Board of Governors of the Federal Reserve System, "Guidance on model risk management," SR Letter 11-7, April 4, 2011 (superseded by SR 26-2).
[22]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[23]. R. Koenker and G. Bassett, "Regression quantiles," Econometrica, vol. 46, no. 1, pp. 33–50, 1978.
[24]. Consumer Financial Protection Bureau, "Fair Lending Report of the Consumer Financial Protection Bureau," annual report, 2023.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Evolution in Remote Sensing Change Detection: Methods, Challenges, and Next AI Perspectives

Zeynep Açikgöz
Computer Engineering
Düzce University
zeynepacikgoz@duzce.edu.tr
Düzce, Türkiye
ORCID:0009-0003-9032-3144

İbrahim Alper Doğru
Computer Engineering
Gazi University
iadogru@gazi.edu.tr
Ankara, Türkiye
ORCID:0000-0001-9324-7157

Kenan Toklu
Civil Engineering
Namık Kemal University
ktoklu@nku.edu.tr
Tekirdağ, Türkiye
ORCID:0000-0003-1288-845X

Abstract
Remote sensing change detection (CD) has undergone a dramatic transformation over the past decade, evolving from classical pixel-difference methods to sophisticated deep learning architectures capable of semantic understanding. This study presents a comprehensive review of advances in remote sensing CD from 2020 to 2026, covering the full task taxonomy, Binary Change Detection (BCD), Multi-Class Change Detection (MCD), Semantic Change Detection (SCD), Change Captioning (CC), and Change Detection Data Synthesis (CDDS), as well as the architectural progression from CNN-based Siamese networks to Transformer and Mamba state space model frameworks. We further examine the frontier of next-generation AI perspectives, including Vision Foundation Models (SAM, CLIP, DINO), Vision-Language Models, and autonomous CD Agent systems that integrate large language models with visual encoders for reasoning-based geospatial change analysis. Representative benchmark datasets (LEVIR-CD, WHU-CD, SECOND, xBD, LEVIR-CC), evaluation metrics, and key open challenges, pseudo-changes, label scarcity, domain gap, and multi-class complexity, are systematically discussed. Finally, we highlight forward-looking developments around multimodal integration, data synthesis, and lightweight architectures for edge execution.

Keywords: Change Detection (CD), Remote Sensing, Deep Learning (DL), Earth Observation (EO), Satellite Imagery

I. INTRODUCTION

Remote sensing change detection (RSCD) is the foundational discipline for identifying and localizing physical changes on the Earth's surface by analyzing multi-temporal imagery acquired over the same geographic region at different times [1]. Mathematically, given a pair of co-registered images I_1 and I_2 taken at distinct time points $t_1 < t_2$, a CD algorithm generates a change map in which each spatial location is assigned a label ranging from binary indicators to sophisticated semantic characterizations [2]. As the frequency and resolution of satellite revisits continue to increase, driven by open-access programs like Copernicus and commercial constellations such as Maxar and Planet, the demand for automated, scalable, and semantically rich CD has transitioned from a research objective to an operational necessity for urban sprawl monitoring, disaster management, and environmental tracking [3]. Historically, the field emerged from visual aerial analysis and evolved into algorithmic approaches relying on pixel-level algebraic operations, such as image differencing, Change Vector Analysis (CVA), and Principal Component Analysis (PCA) [4]. Although these classical methods were interpretable and computationally efficient, they were fundamentally limited by handcrafted features that could not capture the complex spatial context needed to distinguish meaningful surface changes from environmental noise [5]. This groundbreaking shift toward deep learning, which gained momentum after 2015 with the emergence of Fully Convolutional Networks (FCNs) and Siamese architectures, ushered in a new era of performance innovation. These deep learning frameworks significantly elevated benchmark F1 scores by

learning hierarchical representations and establishing the dual-branch paradigm as the dominant approach in the field. As illustrated in Figure 1, the conceptualization of these methodologies is often categorized by their spatial granularity, where pixel-level analysis provides high-resolution precision for simple applications, while object-level approaches utilize texture and shape to enhance noise resistance; ultimately, scene-level methods offer a holistic view of macroscopic regional trends, ensuring scalability on a global scale. Since 2020, RSCD has entered an accelerating period of development, moving from convolutional maturity toward advanced Transformer architectures and the current frontier of Next AI [6].

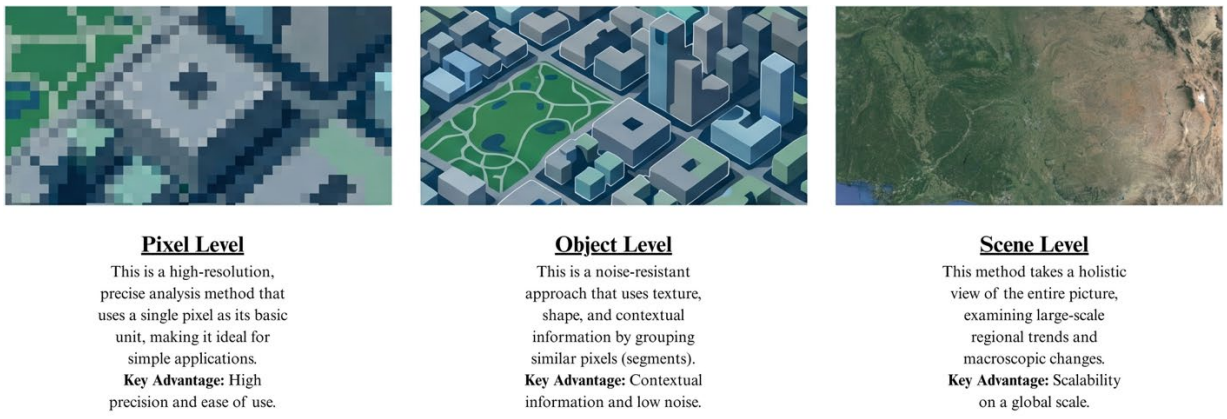


Figure 1. Classification of RSCD methodologies based on spatial granularity.

This study provides a comprehensive survey and synthesis of the field's evolution from 2020 to 2026, a timeframe marked by the release of the LEVIR-CD benchmark and the rise of high-performing Siamese models. We present a structured review of the full task taxonomy, including Binary, Multi-Class, Semantic, and Change Captioning, while positioning emerging paradigms such as Mamba state-space models, Vision Foundation Models like SAM and DINO, and autonomous CD Agent systems within a coherent framework. By cataloging major benchmark datasets and evaluating current performance standards, this work identifies persistent challenges such as pseudo-changes and domain gaps, ultimately outlining forward-looking research directions centered on multimodal integration and self-supervised learning for the next generation of intelligent Earth observation.

II. TASK TAXONOMY AND CONCEPTUAL FRAMEWORK

The organizational structure of modern Remote Sensing Change Detection (RSCD) has evolved into a multi-layered hierarchy that transcends simple pixel-wise comparisons [7]. To systematically evaluate the field's progress, CD tasks are organized along two orthogonal axes: the granularity of the output, ranging from pixel-level masks to high-level scene descriptions, and the semantic richness of the change representation, which dictates the level of categorical or linguistic detail provided [8]. As illustrated in Figure 2, the current research landscape is defined by how raw multi-temporal signals are transformed into actionable geospatial intelligence, categorized not only by the nature of the change output but also by the data source configurations. From the foundational classification of "change vs. no-change" to the generative synthesis of entirely new training samples, this taxonomy covers the field's primary dimensions.

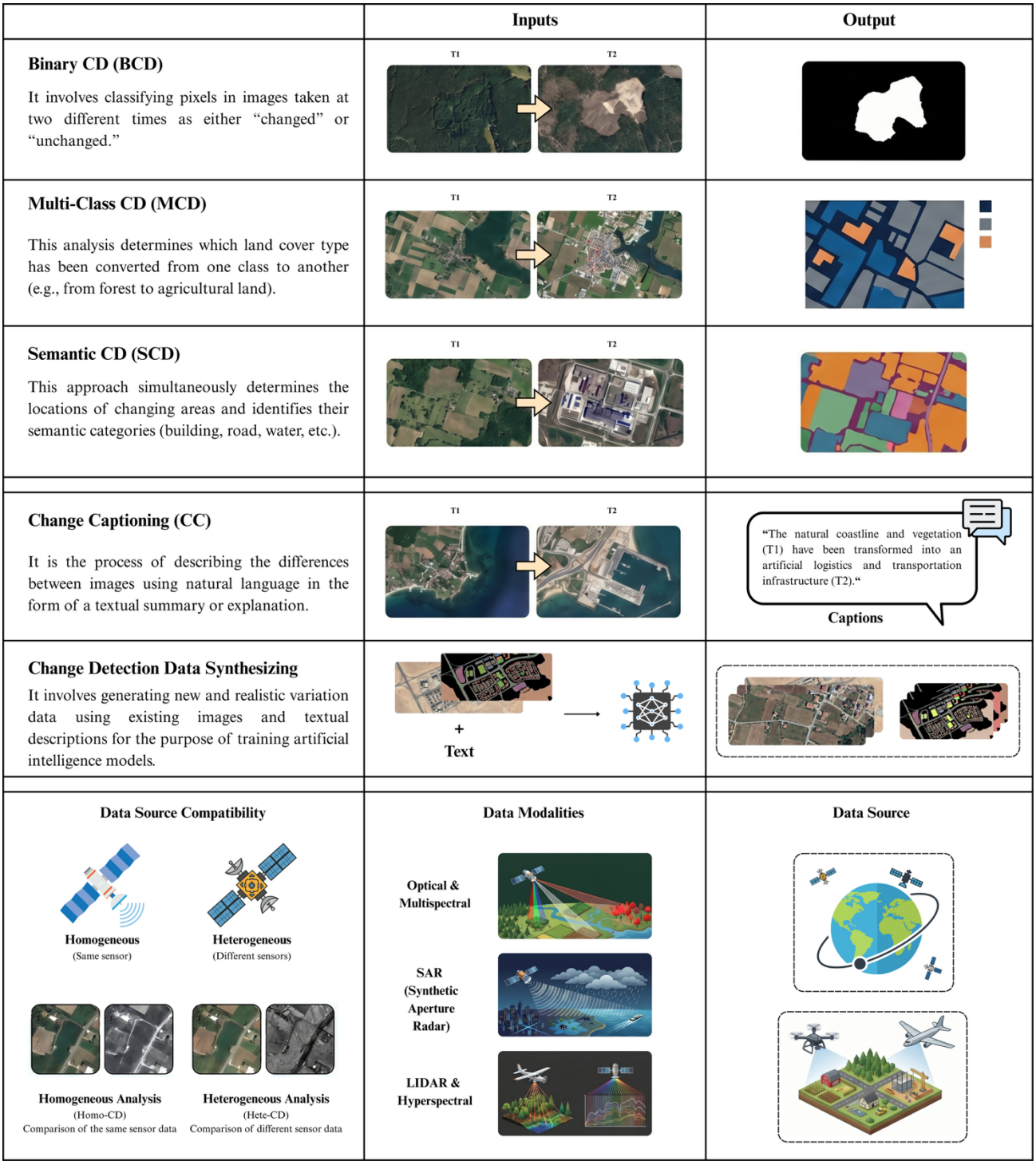


Figure 2. Comprehensive overview of the RSCD conceptual framework.

The contemporary landscape of Remote Sensing Change Detection (RSCD) is defined by a sophisticated task taxonomy that organizes methodologies along two primary axes: the granularity of the spatial output and the semantic richness of the change representation. At the foundational level, Binary Change Detection (BCD)

remains the most extensively researched area, focusing on producing a single-channel mask to distinguish "changed" from "unchanged" pixels [9]. While BCD provides the structural basis for specific applications, such as building footprint extraction or disaster localization, it faces the persistent challenge of filtering out pseudo-changes, such as seasonal vegetation cycles and illumination variations, that do not represent meaningful land-cover transitions. Dominant benchmarks such as LEVIR-CD and WHU-CD remain the gold standard for evaluating BCD performance in high-resolution environments [10]. Advancing beyond simple identification, Multi-Class Change Detection (MCD) categorizes detected differences into predefined semantic labels, such as urban expansion or water body gain [11]. This approach, exemplified by the xBD dataset for grading disaster damage, offers more actionable insights but introduces significant complexity due to inter-class confusion [12]. Taking semantic granularity to its peak, Semantic Change Detection (SCD), or "from-to" analysis to predict both pre-change and post-change land-use classes jointly [13]. This dual-labeling strategy enables comprehensive characterization of geographic transitions and effectively models the precise nature of Earth's surface evolution. The field has recently expanded into multimodal and generative frontiers to bridge the gap between raw data and human reasoning. Change Captioning (CC) leverages vision-language models to generate natural-language descriptions of observed dynamics, as demonstrated by the LEVIR-CC benchmark, transforming pixel-level detections into interpretable reports [14]. Simultaneously, Change Detection Data Synthesis (CDDS) addresses the chronic shortage of annotated samples by employing GANs and modern Diffusion Networks to generate photorealistic bi-temporal pairs with consistent labels [15]. These specialized tasks are further influenced by a diverse range of input modalities, from sub-meter VHR optical imagery and multispectral Sentinel data to all-weather Synthetic Aperture Radar (SAR), which collectively form a complex modality-task matrix that defines the current era of intelligent geospatial analysis.

III. EVOLUTION OF DEEP LEARNING ARCHITECTURES

The technological trajectory of Remote Sensing Change Detection has progressed through distinct stages of innovation, shifting from localized feature extraction to linear-complexity global modeling [16]. The foundations of modern deep learning in this domain were established through the Siamese network paradigm and U-Net variants between 2018 and 2021. By utilizing two weight-sharing branches to process bi-temporal images in parallel, researchers developed three canonical fusion strategies: early fusion (FC-EF), mid-fusion with feature concatenation (FC-Siam-conc), and mid-fusion with feature differencing (FC-Siam-diff) [17]. These architectures relied heavily on the U-Net's skip connections to preserve spatial detail in high-resolution scenes. A pivotal advancement during this era was the BIT model, which successfully bridged CNN feature extraction with Transformer-based context modeling. By using compact token representations to encode spatial-temporal dependencies, BIT achieved state-of-the-art results on the LEVIR-CD and WHU-CD benchmarks while maintaining manageable model complexity [18]. As the field moved into 2022 and 2023, the success of Vision Transformers (ViTs) led to a revolution in architectural design. Models such as ChangeFormer replaced traditional convolutions with a fully Transformer-based Siamese structure, utilizing hierarchical encoders and MLP decoders to demonstrate that long-range global context is superior for change boundary detection [19]. Other variations, such as SwinSUNet, integrated shifted-window self-attention into U-shaped architectures to enable more efficient multi-scale feature extraction, while TransUNetCD employed a hybrid CNN-Transformer encoder to reduce false alarms [20]. The current frontier, beginning in 2024, is defined by the emergence of Mamba and Structured State Space Models (SSMs). These architectures offer a compelling solution to the Transformer's limitations by providing linear computational complexity while preserving global modeling capacity. ChangeMamba represents the first systematic application of this technology to change detection, introducing specialized spatio-temporal mechanisms such as Serial, Parallel, and Cross-Temporal State Space learning [21]. Subsequent innovations, such as RS-Mamba, further refined this approach by introducing omnidirectional selective-scan strategies better suited to the nondirectional nature of remote sensing imagery [22]. Table 1 presents a curated selection of representative DL-based CD methods

organized by architectural family, highlighting key contributions, benchmark datasets evaluated, and publication year.

Table 1. Landmark studies in remote sensing change detection (2021–2026)

Reference & Year	Core Architecture	Task Type	Key Innovation / Contribution	Datasets Used	Performance Metric
SNUNet-CD (Fang et al., 2021/2022)	Densely Connected CNN	BCD	**Dense Feature Fusion:** Mitigates localization bias in deep layers, preserving fine-grained spatial details.	LEVIR-CD, WHU-CD, DSIFN-CD	WHU-CD F1: 91.59%
ChangeFormer (Bandara & Patel, 2022)	Pure Vision Transformer	BCD, MCD	**CNN-Free Design:** Eliminates convolutions, relying purely on hierarchical multi-scale self-attention.	LEVIR-CD+, WHU-CD, DSIFN-CD	LEVIR-CD+ F1: 94.15%
ChangeMask (Zheng et al., 2022)	Multi-Task Encoder-Decoder	SCD	**Temporal Symmetry:** Models semantic-change causal relationships through a temporally symmetric transformer.	SECOND	SECOND mIoU: 80.61%
ChangeCLIP (Dong et al., 2024)	Vision-Language Model (VLM)	BCD, Open-Vocabulary	**Cross-Modal Alignment:** Leverages text prompts for zero-shot detection, bypassing the need for costly pixel annotations.	LEVIR-CD, WHU-CD	SOTA in Zero-Shot / Prompting
ChangeMamba (Chen et al., 2024)	State Space Model (Mamba)	BCD, SCD	**Linear Complexity:** Breaks the quadratic computational bottleneck of Transformers via spatiotemporal state space.	LEVIR-CD+, WHU-CD, SECOND	WHU-CD F1: 95.71%
Change-Agent (Liu et al., 2024)	LLM-Driven Autonomous Agent	Interactive CD / CC	**Interactive Interpretation:** Fuses pixel features with LLM semantics for conversational reasoning and visual QA.	LEVIR-CC, Custom datasets	High Semantic Fidelity

Changen2 (Zheng et al., 2025)	Generative Foundation Model	CDDS (Data Synthesis)	**Solving Data Scarcity:** Synthesizes highly realistic bi-temporal images via diffusion to overcome annotation bottlenecks.	Synthetic LEVIR / WHU	Radical downstream F1 boost

*(**BCD** = Binary Change Detection, **MCD** = Multi-class Change Detection, **SCD** = Semantic Change Detection, **CC** = Change Captioning, **CDDS** = Change Detection Data Synthesis).*

IV. FROM UNIMODAL TO MULTIMODAL CHANGE DETECTION

Most early DL-based CD methods assumed co-registered optical image pairs from the same sensor. Real-world operational deployment, however, demands robust performance across heterogeneous data sources [23]. Synthetic Aperture Radar (SAR, e.g., Sentinel-1) is weather-independent and day/night-capable, making it invaluable for disaster response, where optical imagery is frequently cloud-contaminated [24]. However, SAR's coherent speckle noise and fundamentally different imaging physics, backscatter intensity versus spectral reflectance, make direct bi-temporal comparison nontrivial. Heterogeneous CD (HetCD) methods address this by aligning optical and SAR representations within a shared latent feature space before change analysis, often via domain adaptation or cross-modal attention [25]. Hyperspectral imagery (HSI) provides hundreds of narrow spectral bands, enabling fine-grained material discrimination but introducing the curse of dimensionality [26]. Recent approaches apply band selection and joint spectral-spatial frequency-domain attention to exploit hyperspectral richness while controlling computational cost.

V. BENCHMARK DATASET

Table 2 provides a comprehensive synthesis of the most widely adopted open-source change detection datasets, spanning a diverse range of task paradigms including Binary Change Detection (BCD), Multi-Class Change Detection (MCD), and Semantic Change Detection (SCD). By systematically detailing critical metadata such as spatial resolution, total number of image pairs, geographic distribution, and specific land-cover categories, this summary serves as a foundational reference for the research community. It is designed to facilitate rigorous, fair performance comparisons across varying architectural frameworks and to promote further methodological developments. Furthermore, the inclusion of specialized benchmarks for disaster assessment and urban evolution enables researchers to select the most appropriate data environments to test the robustness and generalizability of "Next AI" models, such as Mamba-based and Foundation-driven systems, in real-world geospatial scenarios [27].

Table 2. Summary of Representative Remote Sensing Change Detection Datasets

Dataset	Year	Task	GSD	Size	Patch Size	Description
LEVIR-CD	2020	BCD	0.5 m	637 pairs	1024×1024	High-resolution Google Earth images from Texas, USA, focusing on building growth and decline.
WHU-CD	2018	BCD	0.2–0.3 m	1 pair (32,507×15,354)	256×256	Aerial images of Christchurch, New

						Zealand, capturing building changes after a 6.3-magnitude earthquake.
xBD	2019	MCD / BDA	0.3–0.8 m	22,068 images	1024×1024	Global disaster dataset assessing building damage across four severity levels for various natural disasters.
SECOND	2021	SCD	0.5–3 m	4,662 pairs	512×512	Aerial images from Chinese cities annotating 6 land-cover classes and 30 semantic change transitions.
CDD	2018	BCD	0.03–1 m	16,000 patches	256×256	Google Earth images capture seasonal variations in natural and artificial features, such as snow and leaves.
HRSCD	2019	SCD	0.5 m	291 pairs (10k×10k)	Various	Large-scale RGB aerial images from France, annotating 5 land-cover classes for pixel-level semantic changes.
LEVIR-CC	2022	CC (Captioning)	0.5 m	10,077 pairs	256×256	Image pairs, each with 5 natural language descriptions, covering various urban and geographical changes.
MineNetCD	2025	BCD	1.2 m	~70,000 patches	256×256	Comprehensive global dataset featuring high-resolution images from 100 open-pit mining sites across six continents.

*(**BCD** = Binary Change Detection, **MCD** = Multi-class Change Detection, **SCD** = Semantic Change Detection, **BDA** = Building Damage Assessment, **CC** = Change Captioning.)*

VI. NEXT AI PERSPECTIVES

The most transformative development in post-2022 remote sensing CD is the integration of large-scale pre-trained foundation models. These models, trained on billions of images or image-text pairs, exhibit remarkable zero-shot and few-shot generalization, directly addressing the perennial data scarcity problem in CD.

1. Vision Foundation Models: SAM, CLIP, and DINO

Meta AI's Segment Anything Model (SAM) demonstrated unprecedented zero-shot segmentation capability via a promptable architecture [28]. SAM-CD adapted SAM for remote sensing CD by injecting temporal-difference features into SAM's prompt encoder and adding lightweight change-specific adapters, transforming

a general segmentation model into a change-aware detector without full retraining [29]. ChangeCLIP [30] leveraged CLIP's contrastive language-image pre-training by formulating CD as a visual-semantic matching task: change regions are identified by measuring cosine similarity between bi-temporal image embeddings and text prompts describing change semantics. DINO-based self-supervised visual features have been employed to improve representation quality in low-label regimes [31].

2. Vision-Language Models and LLMs

The fusion of vision encoders with Large Language Models (LLMs) has opened an entirely new dimension for CD. Remote Sensing Visual Question Answering (RSVQA) frameworks enable analysts to query change maps via natural-language dialogue. RS-LLaVA [32] fine-tuned the LLaVA multimodal architecture on remote-sensing instruction-following datasets, enabling dialogue such as 'How many new structures appeared in the industrial zone after the flood?' The LEVIR-CC Change Captioning benchmark [33] specifically evaluates this paradigm, requiring models to produce descriptive sentences rather than binary masks, a fundamentally different evaluation philosophy that aligns model outputs with human analytical workflows.

3. Change Detection Agents

The most frontier concept is the emergence of autonomous CD Agent systems [34]. Rather than passive detectors, these systems combine LLM-based reasoning engines with visual encoders and external tool APIs to execute multi-step analytical workflows: automatically selecting relevant time-step pairs, choosing appropriate detection sub-models, performing spatial reasoning over change maps, cross-referencing with auxiliary geospatial databases, and producing structured decision reports [35]. Such agentic pipelines draw on frameworks such as Toolformer [36] and represent a paradigm shift from task-specific models toward general-purpose remote sensing intelligence systems capable of dynamic planning and causal inference.

VII. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The field of Remote Sensing Change Detection (RSCD) faces enduring obstacles, primarily driven by pixel-level pseudo-changes caused by seasonal vegetation shifts, illumination variations, and cloud shadows that mimic real land-cover conversions [37]. This issue is exacerbated by severe domain gaps, causing models optimized for distinct sensors or geographic zones to undergo substantial performance decay when deployed in unseen environments. Compounding these technical challenges are chronic data scarcity and critical class imbalances, as high-fidelity, dual-temporal annotations demand intensive human expert labor. At the same time, true change events comprise only a nominal fraction of geographic scenes [38]. To circumvent these limitations, the research community is pivoting toward self-supervised and weakly supervised paradigms, leveraging techniques such as masked autoencoders to extract robust representations from vast unlabeled satellite repositories. In tandem, controllable generative data synthesis via advanced Diffusion Models is emerging as a means to artificially engineer photorealistic multi-temporal image pairs, thereby alleviating label bottlenecks and addressing scarce training scenarios [39]. Concurrently, the transition to complex multi-class and semantic change detection tasks calls for advanced multi-task learning frameworks capable of parsing intricate, visually synchronous land transitions. To bridge the gap toward practical deployment, considerable efforts are also devoted to compressing deep models through knowledge distillation and quantization for edge-hardware execution on resource-constrained UAVs and nanosatellites [40]. Ultimately, the overarching frontier in Earth observation lies in fusing these sensor models with Large Language Models to establish multimodal autonomous agents that can dynamically integrate ancillary geospatial layers and deliver complete, reasoning-based decision reports.

VIII. CONCLUSION

This comprehensive survey traces the architectural trajectory of deep learning-based remote sensing change detection from 2020 to 2026, illustrating how the paradigm shifted from CNN-based Siamese networks to Transformers and linear-complexity Mamba architectures. Beyond structural evolution, the work systematically categorizes a multi-tiered task taxonomy that encompasses binary, multi-class, semantic, captioning, and generative synthesis dimensions. It evaluates eleven pivotal benchmark datasets and representative methodologies. It highlights four critical frontiers for ongoing methodological refinement, specifically targeting the persistent bottlenecks of pseudo-changes, labeling constraints, domain gaps, and complex semantic transitions. Furthermore, the analysis emphasizes how the integration of Vision Foundation Models, Vision-Language frameworks, and autonomous agentic systems is fundamentally repositioning change detection as a core pillar of generalized geospatial intelligence. Ultimately, the study underscores a vital imperative for the research community to navigate this technological transition without compromising domain-specific operational rigor, emphasizing the critical need for sensor-agnostic robustness, calibrated uncertainty estimation, and highly interpretable analytical outputs.

REFERENCES

[1]. Q. Zhu, X. Guo, Z. Li, and D. Li, 'A review of multi-class change detection for satellite remote sensing imagery', Geo-spatial Information Science, vol. 27, no. 1, pp. 1–15, Jan. 2024, doi: 10.1080/10095020.2022.2128902.

[2]. Okorie, A. (2022). Image Segmentation-Registration Cooperative Techniques Applied to Biomedicine and Remote Sensing (Doctoral dissertation, Delaware State University).

[3]. A. Duggan, B. Andrade, and H. Afli, 'Advancing Earth observation: a survey on AI-powered image processing in satellites', European Journal of Remote Sensing, vol. 58, no. 1, p. 2567921, Dec. 2025, doi: 10.1080/22797254.2025.2567921.

[4]. Bi-temporal change detection for topographic map updates using panoptic segmentation of VNIR orthophotos and LiDAR data | GeoInformatica | Springer Nature Link'. Accessed: May 22, 2026. [Online]. Available: https://link.springer.com/article/10.1007/s10707-025-00561-z

[5]. M. A. El-Omairi and A. E. Garouani, 'Satellite-based lithological Mapping: A review of Sensors, Methods, and emerging Machine-Learning approaches', International Journal of Applied Earth Observation and Geoinformation, vol. 149, p. 105309, May 2026, doi: 10.1016/j.jag.2026.105309.

[6]. Transformers in Remote Sensing: A Survey'. Accessed: May 21, 2026. [Online]. Available: https://www.mdpi.com/2072-4292/15/7/1860

[7]. Y. Chen et al., 'Deep Learning for High-Resolution Remote Sensing Change Detection: Task Taxonomy, Datasets, and Perspectives', IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 27886–27904, 2025, doi: 10.1109/JSTARS.2025.3622204.

[8]. L. Wang, M. Zhang, X. Gao, and W. Shi, 'Advances and Challenges in Deep Learning-Based Change Detection for Remote Sensing Images: A Review through Various Learning Paradigms', Remote Sensing, vol. 16, no. 5, p. 804, Jan. 2024, doi: 10.3390/rs16050804.

[9]. M. M. Elisy, X. Xiao, M. H. Sharouda, J. Gong, and D. Li, 'ReSCD-Net: Resolution-Aware Deep Learning for Building Change Detection in High-Resolution Satellite Imagery', IEEE Transactions on Geoscience and Remote Sensing, pp. 1–1, 2026, doi: 10.1109/TGRS.2026.3662471.

[10]. P. Chen, T. Lu, X. Li, S. Li, Y. Wu, and J. Xiao, 'A Multi-Temporal Remote Sensing Image Dataset for Building Change Detection in Chinese Regions', in IGARSS 2024 - 2024 IEEE International

Geoscience and Remote Sensing Symposium, Jul. 2024, pp. 10044–10047. doi: 10.1109/IGARSS53475.2024.10642794.

[11]. Y. Zhou, J. Wang, J. Ding, B. Liu, N. Weng, and H. Xiao, 'SIGNet: A Siamese Graph Convolutional Network for Multi-Class Urban Change Detection', Remote Sensing 2023, Vol. 15, Page 2464, vol. 15, no. 9, p. 2464, May 2023, doi: 10.3390/RS15092464.

[12]. G. Simantiris, K. Bacharidis, A. Papanikolaou, P. Giannakakis, and C. Panagiotakis, 'AIFloodSense: A Global Aerial Imagery Dataset for Semantic Segmentation and Understanding of Flooded Environments', Remote Sensing, vol. 18, no. 6, p. 938, Jan. 2026, doi: 10.3390/rs18060938.

[13]. Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, 'Single-Temporal Supervised Learning for Universal Remote Sensing Change Detection', Int J Comput Vis, vol. 132, no. 12, pp. 5582–5602, Dec. 2024, doi: 10.1007/s11263-024-02141-4.

[14]. S. Zou, Y. Wei, Y. Xie, M. Lao, and X. Luan, 'Remote Sensing Image Change Captioning: A Comprehensive Review', Int J Multimed Info Retr, vol. 14, no. 3, p. 26, Jul. 2025, doi: 10.1007/s13735-025-00375-7.

[15]. D. Peng, M. Liu, Y. Zhang, and H. Guan, 'Toward Label-Efficient Deep Learning Change Detection for Remote Sensing Imagery: A Comprehensive Review', The Photogrammetric Record, vol. 40, no. 191, p. e70021, 2025, doi: 10.1111/phor. 70021.

[16]. Y. Shang et al., 'Artificial Intelligence for Remote Sensing: Progress, Challenges, and Perspectives', IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 1–36, Jan. 2026, doi: 10.1109/JSTARS.2026.3656185.

[17]. R. Ashraf, A. Priadana, X.-T. Vo, G. Cao, T.-D. Tran, and K.-H. Jo, 'ESIFCD: A Dual-Temporal Cross-Domain Fusion Network for Remote Sensing Change Detection', IEEE Access, vol. 13, pp. 167227–167249, 2025, doi: 10.1109/ACCESS.2025.3612956.

[18]. D. Zhang, F. Wang, L. Ning, Z. Zhao, J. Gao, and X. Li, 'Integrating SAM With Feature Interaction for Remote Sensing Change Detection', IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–11, 2024, doi: 10.1109/TGRS.2024.3483775.

[19]. W. G. C. Bandara and V. M. Patel, 'A Transformer-Based Siamese Network for Change Detection', in IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Jul. 2022, pp. 207–210. doi: 10.1109/IGARSS46834.2022.9883686.

[20]. F. Yang et al., 'ConvFormer-CD: Hybrid CNN–Transformer With Temporal Attention for Detecting Changes in Remote Sensing Imagery', IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1–15, 2025, doi: 10.1109/TGRS.2025.3544651.

[21]. H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, 'ChangeMamba: Remote Sensing Change Detection With Spatiotemporal State Space Model', IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–20, 2024, doi: 10.1109/TGRS.2024.3417253.

[22]. S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, and W. Ouyang, 'RS-Mamba for Large Remote Sensing Image Dense Prediction', IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–14, 2024, doi: 10.1109/TGRS.2024.3425540.

[23]. X. X. Zhu et al., 'Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources', IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8–36, Dec. 2017, doi: 10.1109/MGRS.2017.2762307.

[24]. M. Seo, J. Jung, and D.-G. Choi, 'Improved Flood Insights: Diffusion-Based SAR-to-EO Image Translation', Remote Sensing, vol. 17, no. 13, p. 2260, Jan. 2025, doi: 10.3390/rs17132260.

[25]. L. T. Luppino et al., 'Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images', IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 1, pp. 60–72, Jan. 2024, doi: 10.1109/TNNLS.2022.3172183.

[26]. S. Sudharsan, R. Hemalatha, and S. Radha, 'Deep learning techniques for hyperspectral imaging classification: challenges and opportunities', Environ Earth Sci, vol. 85, no. 9, p. 223, Apr. 2026, doi: 10.1007/s12665-026-12942-0.

[27]. M. Ahmad et al., 'A comprehensive survey for Hyperspectral Image Classification: The evolution from conventional to transformers and Mamba models', Neurocomputing, vol. 644, p. 130428, Sep. 2025, doi: 10.1016/j.neucom.2025.130428.

[28]. A. Kirillov et al., 'Segment Anything', presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026. Accessed: May 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html

[29]. L. Ding, K. Zhu, D. Peng, H. Tang, K. Yang, and L. Bruzzone, 'Adapting Segment Anything Model for Change Detection in VHR Remote Sensing Images', IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–11, 2024, doi: 10.1109/TGRS.2024.3368168.

[30]. S. Dong, L. Wang, B. Du, and X. Meng, 'ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning', ISPRS Journal of Photogrammetry and Remote Sensing, vol. 208, pp. 53–69, Feb. 2024, doi: 10.1016/j.isprsjprs.2024.01.004.

[31]. M. Caron et al., 'Emerging Properties in Self-Supervised Vision Transformers', presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660. Accessed: May 21, 2026. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper

[32]. Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, 'RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery', Remote Sensing, vol. 16, no. 9, p. 1477, Jan. 2024, doi: 10.3390/rs16091477.

[33]. C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, 'Remote Sensing Image Change Captioning With Dual-Branch Transformers: A New Method and a Large Scale Dataset', IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–20, 2022, doi: 10.1109/TGRS.2022.3218921.

[34]. C. Liu, K. Chen, H. Zhang, Z. Qi, Z. Zou, and Z. Shi, 'Change-Agent: Toward Interactive Comprehensive Remote Sensing Change Interpretation and Analysis', IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3425815.

[35]. Y. Xu et al., 'Integrating Large Vision Models into Agentic AI Communications and Networking: A Survey', TechRxiv, vol. 2025, no. 1229, doi: 10.36227/techrxiv. 176704888.89434041/v1.

[36]. T. Schick et al., 'Toolformer: Language Models Can Teach Themselves to Use Tools', Advances in Neural Information Processing Systems, vol. 36, pp. 68539–68551, Dec. 2023.

[37]. D. Peng, M. Liu, Y. Zhang, and H. Guan, 'Toward Label-Efficient Deep Learning Change Detection for Remote Sensing Imagery: A Comprehensive Review', The Photogrammetric Record, vol. 40, no. 191, p. e70021, 2025, doi: 10.1111/phor. 70021.

[38]. J. Zhong, S. Zhang, T. Ji, and Z. Tian, 'Enhancing Single-Temporal Semantic and Dual-Temporal Change Perception for Remote Sensing Change Detection'.

[39]. Z. Zhao et al., 'ChangeBridge: Spatiotemporal Image Generation with Multimodal Controls for Remote Sensing', arXiv.org. Accessed: May 21, 2026. [Online]. Available: https://arxiv.org/abs/2507.04678v3

[40]. J. Li, Y. Shi, Q. Hong, and Y. Jia, 'A Scale-Aware Multidomain DETR for Small Object Detection in UAV Remote Sensing Imagery', IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1–20, 2025, doi: 10.1109/TGRS.2025.3624765.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Multi-Objective Hybrid Swarm Optimization for Energy-Delay Trade-off in SDN-enabled MEC

Cherhabil Fatima Z.
Computer Science Department
Mustafa Benboulaid University (Batna2)
Batna, Algeria
f.cherhabil@univ-batna2.dz
ORCID: 0000-0002-6937-5092

Bendib Sonia-Sabrina
Computer Science Department
Mustafa Benboulaid University (Batna2)
Batna, Algeria
ss.bendib@univ-batna2.dz
ORCID: 0000-0003-4674-6185

Sedrati Maamar
Computer Science Department
Mustafa Benboulaid University (Batna2)
Batna, Algeria
m.sedrati@univ-batna2.dz
ORCID: 0000-0003-0444-0399

Abstract
Mobile Edge Computing (MEC) mitigates the limitations of cloud-centric architectures for Internet of Things (IoT) devices by bringing processing closer to the data source. However, jointly optimizing task offloading and resource allocation in SDN-enabled MEC remains a challenging multi-objective problem due to conflicting delay and energy objectives. This paper formulates the problem as a constrained multi-objective optimization model. It proposes three hybrid swarm-based approaches that integrate Particle Swarm Optimization (PSO) with the Whale Optimization Algorithm (WOA), Cuckoo Search (CS), and the Grey Wolf Optimizer (GWO). When evaluated using Pareto-based metrics (Hyper-volume, Spread, and Execution Time), the hybrid approaches achieved competitive overall performance compared with standalone algorithms in convergence quality, diversity preservation, and computational efficiency. Specifically, MOPSO_WOA and MOPSO_GWO achieved improved energy-delay trade-offs with reduced execution times, demonstrating their efficacy for real-time SDN-enabled MEC optimization.

Keywords: Mobile Edge Computing, Energy-Delay Trade-off, Pareto Multi-Objective Optimization, Task Offloading, Internet of Things

I. INTRODUCTION

The Internet of Things (IoT) network generally consists of a large number of latency-sensitive and energy-constrained devices. Mobile Edge Computing (MEC) alleviates the burden on centralized cloud infrastructures by enabling computationally intensive tasks to be offloaded from these devices to nearby edge servers [1]. To manage the dynamic and heterogeneous nature of these environments, MEC is paired with Software-Defined Networking (SDN) [2-4]. In this architecture, an SDN controller leverages global network intelligence to orchestrate computation offloading and resource allocation [5]. However, despite the benefits of SDN-enabled MEC, jointly optimizing task offloading and resource allocation remains a challenging multi-objective optimization (MOO) problem due to the conflicting nature of delay and energy consumption. Offloading all tasks to MEC servers and computing locally are both extreme cases in the decision space. Both mechanisms improve in one aspect while failing in the other. Yet, most IoT applications are constrained by both delay and energy [6]. Solving this MOO problem effectively is crucial to realizing the full potential of MEC. While traditional optimization techniques often struggle to capture the complexity of such environments, especially under real-time constraints, swarm-based algorithms have been adapted to address this challenge because they can navigate NP-hard solution spaces without gradient information [7]. Early efforts utilized weighted-sum methods such as Genetic Algorithms (GA) [8], [9], PSO [10-12], Cuckoo Search (CS) [13], Artificial Bee Colony [14], [15], GWO [16], and WOA [17-19]. However, this approach requires pre-set weights and risks overlooking optimal trade-off solutions. Consequently, Pareto-based MOO algorithms have gained significant

attention for solving multi-objective task offloading problems in MEC because they can simultaneously optimize multiple solutions, reducing the risk of falling into local optima. NSGA-II is a frequent choice for optimizing energy and delay [20-23]. Recent advancements include MOWOA [24] and MOGWO [25] for collaborative edge offloading, as well as MOPSO [26], MOEA [27], and PAES [28]. While these standalone studies demonstrate the strength of evolutionary algorithms in MEC, there remains a need to improve convergence speed, solution diversity, and dynamic adaptability. Thus, striking the right balance between exploration and exploitation is the most challenging task in developing any evolutionary or swarm-based algorithm. Unlike conventional sequential hybridization strategies, in this study, we introduce three hybrid multi-objective approaches, MOPSO_WOA, MOPSO_CS, and MOPSO_GWO, which integrate PSO velocity-driven exploitation with complementary position-update mechanisms from WOA, CS, and GWO inside a unified Pareto optimization framework, along with dynamic parameter tuning for adaptive search behavior. The goal is to improve convergence quality, enhance solution diversity, and effectively balance exploration and exploitation in the search space. The primary contributions of this paper are:

- Formulation of a centralized, SDN-enabled MEC task offloading and resource allocation framework as a constrained MOO problem targeting energy and delay minimization.
- Development of three hybrid swarm-intelligence algorithms incorporating dynamic parameter tuning to enhance convergence speed and solution diversity.
- Rigorous Pareto-based evaluation demonstrating competitive Hyper-volume, Spread, and superior computational efficiency of the hybrid models compared with standalone counterparts.

The rest of the paper is organized as follows: Section 2 details the system model and problem formulation. Section 3 presents the proposed algorithms. Section 4 discusses the evaluation results, and Section 5 concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

1. System Model and Communication

We consider an MEC environment comprising a set of IoT devices randomly deployed and connected to edge servers via a base station (BS). The BS hosts an SDN controller that orchestrates task offloading and resource allocation decisions by generating and distributing offloading strategies to IoT devices [29]. Each user generates a computational task *Qi = (Di, Ci, T_i, max)* representing data size, required CPU cycles, and delay tolerance. The decision variables are the offloading proportion vector $\alpha = (\alpha_1, \dots, \alpha_N)$, where $\alpha_i \in [0,1]$, and the server allocation vector $\beta = (\beta_1, \dots, \beta_N)$, indicating the selected edge server (or "0" for local execution).

The model assumes shared-spectrum uplink communication with simultaneous transmissions. Thus, the uplink data rate for device U_i, over a bandwidth *B,* is modeled by the Shannon-Hartley theorem [27]:

$$R_i = B\ log_2\left(1 + \frac{h_i p_i}{\sigma^2 + \sum_{j=1, j\neq i}^{N} h_j p_j}\right) \quad (1)$$

Where p_i is the transmission power, h_i is the channel gain, σ^2 is the channel noise, aand the ter toplam alt limit j eşittir 1d the ter $\sum_{j=1, j\neq i}^{N} h_j p_j$ represents the effects from other IoT devices.

2. Computation Model

The total energy and delay for each task are computed for both local and remote execution, accounting for transmission and execution.

For device U_i with computational capacity f_i and capacitance coefficient κ^{user}, the local execution time and the energy consumption are given as:

$$T_i^{local} = \frac{C_i}{f_i} \tag{2}$$

$$E_i^{local} = \kappa^{user} (f_i)^2 C_i \tag{3}$$

In remote execution, the orchestrator (SDN controller) is responsible for selecting the appropriate server for task execution. This process is divided into three stages:

- Transmitting the task input data from the device to the BS. The corresponding formulas are given as follows:

$$T_i^{Trans} = \frac{D_i}{R_i} \tag{4}$$

$$E_i^{Trans} = p_i T_i^{Trans} = \frac{p_i D_i}{R_i} \tag{5}$$

- The execution of task Q_i in edge servers, with the corresponding time calculated as follows:

$$T_i^{R-exe} = \frac{C_i}{f_{\beta_i}^{serv}} \tag{6}$$

Here, $f_{\beta_i}^{serv}$ represents the capacity of the selected MEC server β_i computing resources allocated to each user. The energy consumed by the server is not included in our objective function, as we focus on user-side energy.

- Downloading the results from the edge server to the device is considered negligible due to the small output size compared to the input data typical of many IoT monitoring and analysis tasks [7]. This assumption is suitable for monitoring, sensing, and inference applications where the output size is significantly smaller than the input size.

The total response time and energy consumption are given by:

$$T^{total} = \sum_{i=1}^{N} \left[(1 - \alpha_i) T_i^{local} + \alpha_i (T_i^{Trans} + T_i^{R-exe}) \right] \tag{7}$$

$$E^{total} = \sum_{i=1}^{N} \left[(1 - \alpha_i) E_i^{local} + \alpha_i (E_i^{Trans}) \right] \tag{8}$$

3. Optimization Problem

The objective is to find the Pareto optimal set for α and β that simultaneously minimizes total energy consumption and execution delay:

$$\underset{(\alpha,\beta)}{MIN}(\, E^{total}, T^{total}) \tag{9}$$

Subject to:

- Valid offloading decision: $\alpha_i \in [0,1], \forall i \in (1,\dots,N)$
- Valid server selection: $\beta_i \in \{0,1,\dots,M\}, \forall i \in (1,\dots,N)$
- Edge capacity limits: $\sum_{i=1}^{N}\left[\, \delta_{j,\beta_i}\alpha_i\, f_{\beta_i}^{serv}\right] \le\ f_{max}^{serv}, \forall j \in (1,\dots,M)$
- Device capacity: $f_i > 0\ , \forall i \in (1,\dots,N)$
- Strict delay constraints: $\left[(1-\alpha_i)T_i^{local} +\ \alpha_i\left(T_i^{Trans} +\ T_i^{Edge}\right)\right] \le\ T_i^{max}, \forall i \in (1,\dots,N)$

Where: deltaalt j ,beta alt i. δ_{j,β_i} is the Kronecker delta function, which is equal to 1 if $j = \beta_i$ (indicating task Q_i is assigned to server j), and 0 otherwise. This problem is proven to be NP-hard due to its combinatorial nature and interdependent constraints [5], motivating the use of hybrid evolutionary algorithms [30].

III. Proposed Task Offloading and Resource Allocation Algorithms

To address the multi-objective problem, we propose integrating PSO with WOA, CS, and GWO. PSO enhances global search ability through velocity-based updates, WOA achieves faster convergence via spiral mechanisms, CS allows occasional high jumps to improve exploration, and GWO extends the search space using social hierarchy and hunting agents. In addition, dynamic parameter tuning is employed to enhance exploitation capability. The algorithmic workflow proceeds iteratively through the following steps.

- **Initialization and Solution Representation.** The process begins by initializing a population of $nPop$ particles and velocities. Each particle X_i c, consists of a continuous component, büyük harf X alt i. . alfa, representing offloading decisions, and a discrete component, büyük harf X alt i. . beta, representing server assignments.
- **Velocity Update.** Each particle's velocity is adjusted by considering the three main PSO components [31]: the influence of its previous velocity($i.$), its personal ebesposition ($pBest_i$), and the global best position ($gBest$). This is expressed as:

$$V_i(t+1) = w.V_i(t) +\ c_1.r_1.(pBest_i - X_i(t)\,) +\ c_2.r_2.(gBest - X_i(t)\,) \tag{10}$$

Where: w is the inertia weight, c_1 is the cognitive factor, c_2 is the social factor, and both (r_1, r_2) are random numbers in $[0,1]$ to introduce stochastic behavior.

- **Position Updates via Hybridization.** The particle's position is then updated using one of the following algorithms:

1- *WOA-inspired (MOPSO_WOA):* Uses a random variable $r \in [0,1]$. If $r\ <\ 0.5$, a shrinking encircling mechanism is applied to intensify exploitation around the current best solution. Otherwise, a spiral updating

mechanism is employed to enhance exploration around the search space. The threshold value 0.5 ensures an equal probability between exploitation and exploration phases, providing a balanced search behavior [32].

2- *CS-inspired (MOPSO_CS):* Applies a Lévy flight perturbation to the continuous component α:

$$\alpha_i^{new} = clip(\alpha_i + Levy(\gamma), 0, 1) \quad (11)$$

For the discrete component, random reassignment is utilized.

3- *GWO-inspired (MOPSO_GWO):* The particle position is updated according to the social hierarchy of the three best candidate solutions (Alpha, Beta, and Delta wolves), enabling cooperative guidance toward promising regions of the search space [16].

Next, the updated velocity is applied to further refine the particle's position:

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (12)$$

- **Dynamic Parameter Tuning.** To avoid static behavior across generations, key PSO parameters are adapted dynamically over the iterations. The inertia weight w decreases to shift the search from exploration to exploitation, while $cc1\ decreases\ and\ c2\ increases\ with\ similar\ linear\ adaptation\ to\ shift\ the$ cus toward global learning.

$$w(t) = w_{max} - \frac{t}{T}(w_{max} - w_{min}) \quad (13)$$

- **Fitness Evaluation and Repository Update.** During every iteration, the fitness of each new particle is evaluated based on the two objectives (Energy and Delay). Non-dominated solutions are stored in a Pareto-optimal repository and pruned using a crowding-distance criterion to maintain diversity.

The computational complexity of the proposed hybrid algorithms is mainly dominated by particle evaluation, Pareto-repository maintenance, and position-update operations, resulting in an approximate complexity of $O(T \times nPop \times N)$where T is the number of iterations and N the number of tasks.

IV. SIMULATION RESULTS

1. Simulation Setup

The simulation considers a scenario with 3 edge servers and 30 IoT users, which represents a moderate-load SDN-enabled MEC scenario frequently adopted in the literature. Task arrivals are assumed to be static during the optimization process, a common assumption in MEC offloading studies that evaluates optimization performance under a fixed workload. Each user generates a single independent task with inpusize $D_i \in [0, 1]MB$ and requiring $C_i \in [0.2, 1.2]GHz$ of CPU cycles to complete. The maximum tolerable delay (T_i^{max}) per task lies between 1 and 5 seconds. The local computing capacity f_i of each IoT device is set as $f_i \in [0.2, 1.0]$GHz, and the effective capacitance constant for user devices (κ^{user}) is fixed at 5 $\times 10^{-27}$. The wireless communication model uses a bandwidth of B= 50 MHz, user transmission power P_i $\in$ [0, 1] mW, and Gaussian noise power σ ^2 set to 110^-10. Regarding edge servers, each user is allocated a computational load between

5 and 8 GHz (f alt sol parantez beta alt i. sağ parantez üssü s e r v). In contrast, themaximum server capacity is capped at 35 GHz (f_{max}^{serv}). The parameter settings were chosen based on commonly adopted configurations in recent MEC task-offloading studies [33], thereby ensuring heterogeneous workloads and realistic resource constraints.

2. Multi-objective Performance Comparison

The proposed algorithms are compared with three baseline-offloading schemes widely used in academia [34], [35] (all local execution, all offloaded tasks, and random offloading), as well as several well-known swarm-based multi-objective approaches, including MOWOA [24], MOPSO [12], MOGWO [25], and MOCS [13]. Direct numerical comparison with these studies is difficult because they employ different MEC scenarios, task models, and parameter settings. Therefore, we compare against re-implemented baseline algorithms under identical conditions (population sizes, iteration limits, and stopping criteria), ensuring a fair evaluation. The performance analysis focuses on widely used MOO metrics, namely Hyper-volume [24], Spread [36], [37], and Execution Time. All reported results correspond to averages over 20 independent runs to reduce stochastic effects and improve statistical reliability. Figure 1 illustrates the Pareto fronts generated by each algorithm, while Table 1 summarizes the average metric values obtained. The average rank is computed by independently ranking each metric and averaging the resulting ranks.

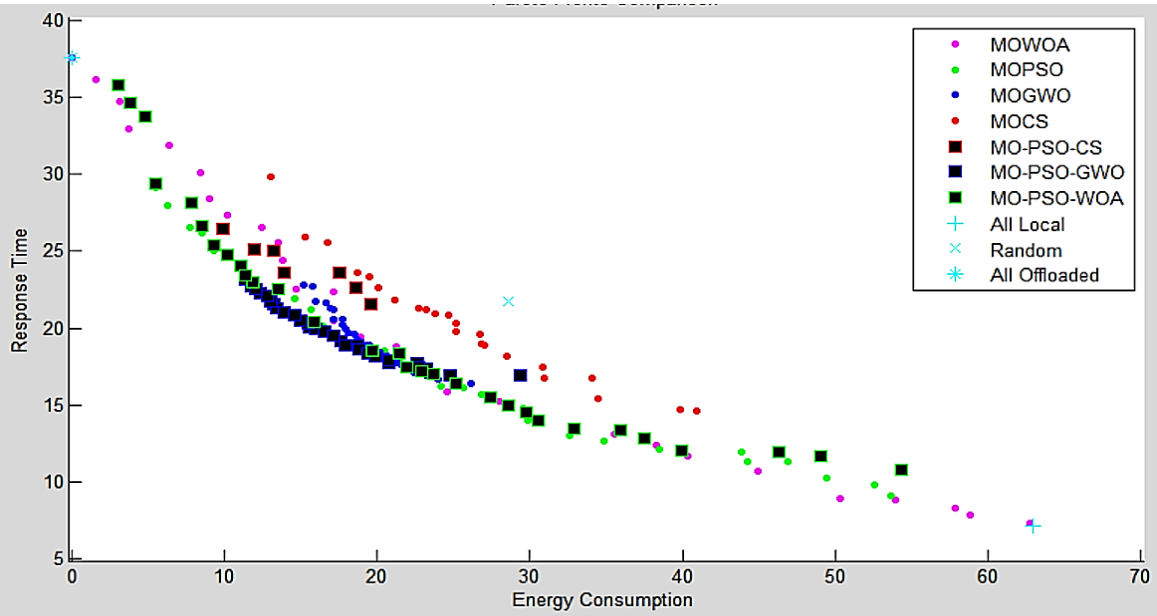


Figure 1. Pareto Fronts Visualization of Each Algorithm

Table 1. Average Performance Metrics across 20 Runs

Algorithm	Hyper-volume ↑	Spread ↓	Execution Time (s) ↓	Average Rank
MOPSO	**1709.64**	0.77	6.69	**2.67**
MOPSO_WOA	1684.28	0.80	3.84	**2.67**
MOPSO_GWO	1482.90	0.88	3.25	3.33
MOWOA	1658.26	**0.71**	11.25	3.67
MOPSO_CS	1133.58	1.01	**2.89**	5.00

MOCS	1342.92	0.91	4.54	5.00
MOGWO	1381.18	0.93	7.42	5.67

The proposed hybrid approaches effectively address the trade-off between energy consumption and delay.

- *Hyper-volume (HV):* While MOPSO achieved the highest raw HV, MOPSO_WOA proved highly competitive (1684.28). MOPSO_WOA and MOPSO_GWO consistently outperformed their standalone variants, indicating that PSO integration successfully limits the aggressive, destabilizing exploration seen in pure WOA and GWO. MOPSO_CS struggled with HV precision due to the large leaps inherent to Lévy flights. Figure 2 shows the HV distribution (average and standard deviation) across algorithms.

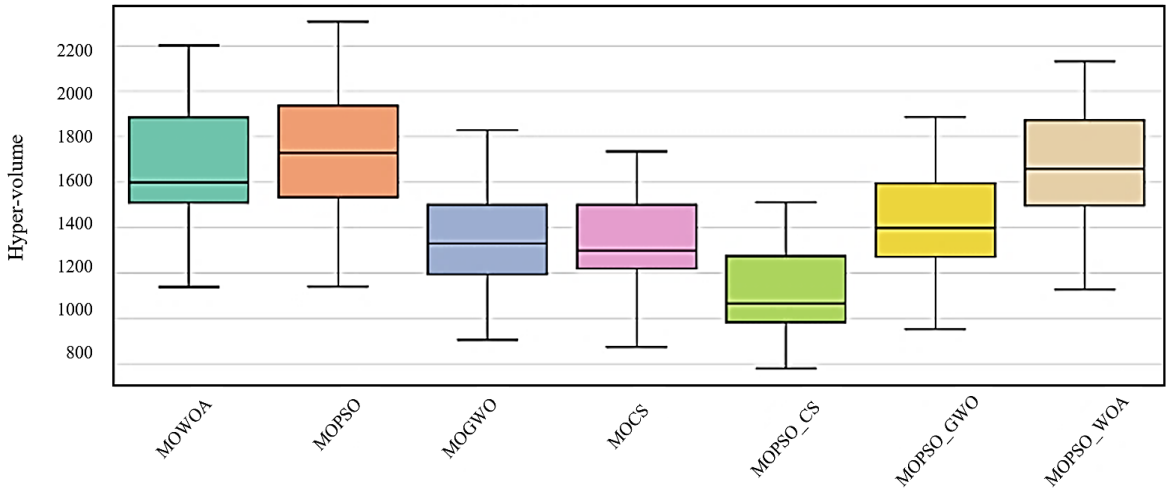


Figure 2. Hyper-volume Distribution across Algorithms

- *Spread (Δ):* MOWOA yielded the best uniformity, but at a massive computational cost. MOPSO_WOA successfully preserved this diversity (0.80) while mitigating the computational overhead. Conversely, MOPSO_CS yielded lower spread values because Lévy-flight perturbations prioritize exploration over Pareto-front regularity. Figure 3 shows the metric distribution across algorithms.

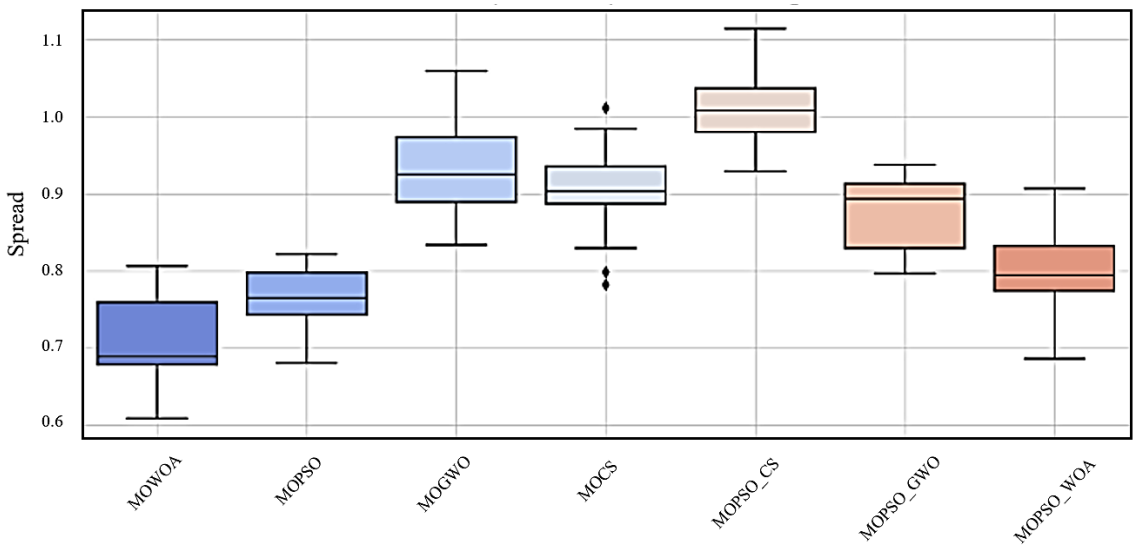


Figure 3. Spread Distribution across Algorithms

- *Execution Time:* This is where the hybrids excel. MOPSO_WOA reduced execution time by roughly 65% compared to MOWOA, converging in just 3.84 seconds. MOPSO_GWO also showed computational economy (3.25s). This confirms that hybridization and dynamic tuning guide the search space efficiently without introducing excessive computational overhead. Figure 4 shows the distribution of this metric across algorithms.

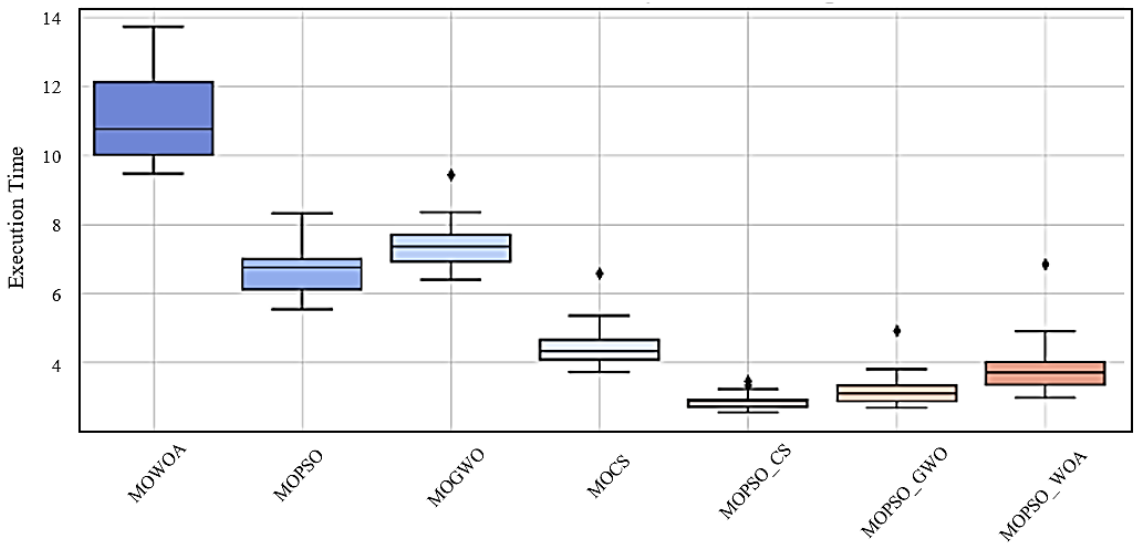


Figure 4. Execution Time Distribution across Algorithms

In summary, the experimental results reveal that no single algorithm dominates all evaluation metrics simultaneously, highlighting the inherent trade-offs between convergence quality, diversity preservation, and computational efficiency in MOO. While MOPSO achieved the highest HV values, the proposed hybrid approaches demonstrated superior balance across all metrics. In particular:

- MOPSO_WOA achieved competitive HV performance while significantly reducing execution time compared to both MOPSO and MOWOA. Although MOPSO and MOPSO_WOA achieved the same average rank (2.67), MOPSO_WOA did so with a substantially lower execution time (3.84 s versus 6.69 s), making it more attractive for latency-sensitive MEC environments.
- MOPSO_GWO provided strong compromise solutions with good convergence behavior and stable computational performance.
- MOPSO_CS prioritized exploration efficiency and execution speed, although at the expense of Pareto-front uniformity.

The obtained results are consistent with recent studies, such as [26], which report that hybridization improves the exploration-exploitation balance by combining complementary search behaviors from multiple swarm intelligence techniques. Consequently, the proposed hybrid methods are better suited for practical SDN-enabled MEC environments where both optimization quality and computational responsiveness are essential.

V. CONCLUSION

This paper introduced three multi-objective hybrid evolutionary algorithms (MOPSO_WOA, MOPSO_CS, and MOPSO_GWO) to optimize the energy-delay trade-off in SDN-enabled MEC environments. By integrating the strengths of PSO with specialized swarm mechanisms, the algorithms effectively balanced exploration and exploitation. Numerical results demonstrated that MOPSO_WOA and MOPSO_GWO provided a favorable compromise between high-quality Pareto convergence, solution diversity, and real-time computational responsiveness. In particular, MOPSO_WOA achieved performance comparable to MOPSO while reducing execution time by more than 40%, making it suitable for latency-sensitive MEC scenarios. Although the evaluation was conducted on a moderate-scale MEC scenario, the proposed framework is independent of network size and can be extended to larger deployments. Future work will investigate scalability under larger MEC infrastructures and physical SDN testbeds, incorporate statistical significance analysis, consider multimedia-oriented applications where downlink transmission may become significant, and compare the proposed swarm-based approaches with recent DRL-based offloading frameworks.

REFERENCES

[1]. L. A. Haibeh, M. C. E. Yagoub, and A. Jarray, "A survey on mobile edge computing infrastructure: Design, resource management, and optimization approaches," *IEEE Access*, vol. 10, pp. 27591–27610, 2022.

[2]. F. Z. Cherhabil, M. Sedrati, and S. S. Bendib, "The integration of software-defined network in mobile edge computing: State of the art," in *20th Int. Conf. Remote Eng. Virtual Instrum.*, Thessaloniki, Greece, 2023.

[3]. P. Hu and W. Chen, "Software-defined edge computing (SDEC): Principles, open system architecture and challenges," in *2019 IEEE SmartWorld*, pp. 8-16, 2019.

[4]. Z. Wang, T. Lv, and Z. Chang, "Computation offloading and resource allocation based on distributed deep learning and software-defined mobile edge computing," Comput. *Netw.*, vol. 205, p. 108732, 2022.

[5]. K. Li, X. Wang, Q. Ni, and M. Huang, "Entropy-based reinforcement learning for computation offloading service in software-defined multi-access edge computing," *Future Gener. Comput. Syst.*, vol. 136, pp. 241–251, 2022.

[6]. L. Li and H. Zhang, "Delay optimization strategy for service cache and task offloading in three-tier architecture mobile edge computing system," *IEEE Access*, vol. 8, pp. 170211–170224, 2020.

[7]. B. Tang, S. Zheng, and Q. Yang, "Joint optimization of energy and delay for computation offloading in vehicular edge computing," *Peer-to-Peer Netw—Appl.*, vol. 16, no. 6, pp. 2681–2695, 2023.

[8]. K. Wang, X. Wang, and X. Liu, "Sustainable Internet of vehicles system: A task offloading strategy based on improved genetic algorithm," *Sustainability*, vol. 15, no. 9, p. 7506, 2023.
[9]. L. Kuang, T. Gong, S. OuYang, H. Gao, and S. Deng, "Offloading decision methods for multiple users with structured tasks in edge computing for smart cities," *Future Gener. Comput. Syst.*, vol. 105, pp. 717–729, 2020.
[10]. L. N. T. Huynh *et al.*, "Efficient computation offloading in multi-tier multi-access edge computing systems: A particle swarm optimization approach," *Appl. Sci.*, vol. 10, no. 1, p. 203, 2019.
[11]. Z. Liu, J. Fan, S. Geng, P. Qin, and X. Zhao, "Joint optimization of task offloading and computing resource allocation in MEC-D2D network," in *2022 IEEE 5th Int. Conf. Comput. Commun. Eng. Technol. (CCET)*, pp. 256-260, 2022.
[12]. B. Cheng, "Multi-population cooperative elite algorithm for efficient computation offloading in mobile edge computing," *J. Grid Comput.*, vol. 21, no. 4, p. 54, 2023.
[13]. M. Alqarni, A. Cherif, and E. Alkayyal, "ODM-BCSA: An offloading decision-making framework based on binary cuckoo search algorithm for mobile edge computing," *Comput. Netw.*, vol. 226, p. 109647, 2023.
[14]. A. B. de Souza *et al.*, "A bee colony-based algorithm for task offloading in vehicular edge computing," *IEEE Syst. J.*, vol. 17, no. 3, pp. 4165–4176, 2023.
[15]. M. Babar *et al.*, "Intelligent computation offloading for IoT applications in scalable edge computing using artificial bee colony optimization," *Complexity*, vol. 2021, no. 1, p. 5563531, 2021.
[16]. A. Abbas, A. Raza, F. Aadil, and M. Maqsood, "Meta-heuristic-based offloading task optimization in mobile edge computing," *Int. J. Distrib. Sens. Netw.*, vol. 17, no. 6, p. 15501477211023021, 2021.
[17]. J. Li, Q. Wang, S. Hu, and L. Li, "Hybrid immune whale differential evolution optimization (HIWDEO) based computation offloading in MEC for IoT," *J. Grid Comput.*, vol. 21, no. 4, p. 70, 2023.
[18]. W. Zhang and W. Zhang, "Research on offloading strategy for mobile edge computing based on improved grey wolf optimization algorithm," *Electronics*, vol. 12, no 11, p. 2533, 2023.
[19]. H. G. T. Pham, Q. V. Pham, A. T. Pham, and C. T. Nguyen, "Joint task offloading and resource management in NOMA-based MEC systems: A swarm intelligence approach," *IEEE Access*, vol. 8, pp. 190463–190474, 2020.
[20]. A. Shahidinejad and M. Ghobaei-Arani, "A metaheuristic-based computation offloading in edge-cloud environment," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 5, pp. 2785–2794, 2022.
[21]. S. Almasri, M. Jarrah, and B. Al-Duwairi, "Multi-objective optimization of task assignment in distributed mobile edge computing," *J. Reliab. Intell. Environ.*, vol. 8, no. 1, pp. 21–33, 2022.
[22]. Y. Ma, X. Li, and J. Li, "An edge computing offload method based on NSGA-II for power Internet of Things," *Internet Things Cloud Comput.*, vol. 9, no. 1, pp. 1–9, 2021.
[23]. Y. Chen, X. Wang, L. Ma, and P. Zhou, "Multi-objective optimization-based task offloading and power control for mobile edge computing," in *Intell. Comput. Theories and Appl.: 17th Int. Conf. (ICIC)*, Shenzhen, China, pp. 3-17, 2021.
[24]. M. Huang, Q. Zhai, Y. Chen, S. Feng, and F. Shu, "Multi-objective whale optimization algorithm for computation offloading optimization in mobile edge computing," *Sensors*, vol. 21, no. 8, p. 2628, 2021.
[25]. N. Nujhat *et al.*, "Task offloading exploiting grey wolf optimization in collaborative edge computing," *J. Cloud Comput.*, vol. 13, no. 1, p. 23, 2024.
[26]. J. Chen *et al.*, "Multi-objective particle swarm optimization algorithm for task allocation and archived guided mutation strategies," *Sci. Rep.*, vol. 15, no. 1, p. 15821, 2025.
[27]. T. Misha, L. Sun, and Z. Chai, "Multi-objective multi-workflow task offloading based on evolutionary optimization," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 37, no. 7, p. 189, 2025.
[28]. Q. Liu, R. Mo, X. Xu, and X. Ma, "Multi-objective resource allocation in mobile edge computing using PAES for Internet of Things," *Wireless Netw.*, pp. 1–13, 2020.
[29]. N. Lo and I. Niang, "SDN-based quality of service architectures in edge-IoT systems: A comprehensive analysis," in *2023 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, pp. 605-611, 2023.

[30]. F. Z. Cherhabil, M. Sedrati, and S. S. Bendib, "AI-based approaches for task offloading, resource allocation and service placement of IoT applications: State of the art," *Int. J. Inf. Commun. Eng.*, vol. 17, no. 2, pp. 137–143, 2023.

[31]. P. Kumar and K. Silambarasan, "Enhancing the performance of healthcare service in IoT and cloud using optimized techniques," *IETE J. Res.*, pp. 1–10, 2019.

[32]. S. Mirjalili and L. Andrew, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, 2016.

[33]. F. Z. Cherhabil, S. S. Bendib, and M. Sedrati, "Hybrid PSO-WOA approach for an efficient task offloading in mobile edge computing," *TELKOMNIKA (Telecommun. Comput. Electron. Control)*, vol. 24, no. 2, pp. 514–526, 2026.

[34]. L. An, Z. Wang, J. Yue, and X. Ma, "Joint task offloading and resource allocation via proximal policy optimization for mobile edge computing network," in *2021 Int. Conf. Netw. Netw. Appl. (NaNA)*, 2021.

[35]. X. Hou *et al.*, "Reliable computation offloading for edge-computing-enabled software-defined IoV," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7097–7111, 2020.

[36]. N. Riquelme, C. Von Lücken, and B. Baran, "Performance metrics in multi-objective optimization," in *2015 Latin American Comput. Conf. (CLEI)*, Arequipa, Peru, pp. 1-11, 2015.

[37]. M. Li and J. Zheng, "Spread assessment for evolutionary multi-objective optimization," in *Evol. Multi-Criterion Optim.*, Berlin, Heidelberg, pp. 216-230, 2009.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Enhancing Small Object Detection in Mammography Using a Patch-Based YOLO Framework

Aray Aryn
Kazakh National Research Technical University K.I. Satpayev
Almaty, Kazakhstan
arynaray1990@gmail.com
ORCID: 0000-0001-7023-0424

Orken Mamyrbayev
Institute of Information and Computing Technologies
Almaty, Kazakhstan
morkenj@mail.ru
ORCID: 0000-0001-8318-3794

Turdybek Kurmetkan
Al-Farabi Kazakh National University
Almaty, Kazakhstan
turdybek.narat86@gmail.com
ORCID: 0000-0003-1178-6400

Abstract
Breast cancer remains a leading global health challenge, with early and accurate detection being critical for reducing mortality. While deep learning models, particularly the YOLO (You Only Look Once) architecture, have shown significant potential in assisting radiologists, their performance often degrades when applied to high-resolution mammography. The primary obstacles include the extremely small relative size of lesions, low contrast, and extreme background dominance, which often lead to target "disappearance" during standard image down-sampling. To address these challenges, this study develops a YOLO-based object detection framework integrated with a tiling (patch-based) preprocessing strategy. By subdividing high-resolution mammograms into smaller overlapping tiles, the proposed approach effectively enlarges the scale of malignant lesions relative to the input window and significantly expands the training dataset. This study proposes a YOLO-based framework for detecting malignant lesions in mammography using a patch-based tiling strategy. High-resolution mammograms are divided into overlapping image patches, increasing the visibility of small pathological regions and improving lesion detection performance.

Keywords: Breast Cancer Detection, YOLO Architecture, Mammography, Tiling, Patch-based Training, Small Object Detection

I. INTRODUCTION

Breast cancer remains a major global health challenge. In 2022 alone, it was responsible for an estimated 670,000 deaths worldwide and affected roughly 2.3 million women, making it the most common cancer among women in 157 of the 185 countries surveyed. The disease knows no borders—it occurs in every country and spans all ages after puberty, with incidence rising sharply in later life. While about half of all cases arise in women who have no identifiable risk factors beyond gender and age, a small but noteworthy proportion (0.5–1 %) also develops in men. The burden of breast cancer is unevenly distributed. In nations with very high Human Development Index (HDI) scores, roughly 1 in 12 women will be diagnosed with breast cancer during her lifetime, and 1 in 71 will die from it. By contrast, in low HDI countries, the lifetime risk of diagnosis drops to 1 in 27, yet mortality rises to 1 in 48, underscoring stark inequities in access to early detection and effective treatment [1]. Early and timely diagnosis, coupled with comprehensive treatment, rehabilitation, and palliative care, is essential to curb this burden and preserve patients' functional well-being. When identified at an early stage—often through routine screening mammography or prompt clinical evaluation of symptoms—breast cancer can be treated effectively with surgery, radiation, and systemic therapies tailored to the tumor's biological subtype. Conversely, delayed diagnosis allows tumors to progress from an in situ (non-life-threatening) state to invasive disease that can metastasize to lymph nodes and distant organs, dramatically reducing survival prospects. Against this backdrop, the present article examines current advances in breast cancer detection and diagnosis, emphasizing the pivotal role of mammographic imaging and emerging artificial intelligence-driven

tools in improving early-stage identification, guiding treatment decisions, and ultimately reducing mortality worldwide [1]. Recent advances in artificial intelligence and deep learning have demonstrated strong potential for assisting radiologists in mammographic image analysis. In particular, convolutional neural networks (CNNs) have shown promising performance in tasks such as lesion classification, segmentation, and object detection. Among these approaches, object detection models aim not only to classify abnormalities but also to localize suspicious regions, which is critical for clinical interpretability and decision support. Nevertheless, the direct application of object detection models to full-resolution mammography images remains challenging due to the extremely large image sizes, severe class imbalance, and the very small relative size of malignant lesions compared to the background tissue [2]. State-of-the-art one-stage detectors, such as YOLO-based architectures, are designed for real-time object detection and have achieved remarkable success in natural image analysis. However, their performance typically degrades when applied to medical images containing small objects embedded in complex backgrounds. In mammography, lesions often occupy less than one percent of the total image area, making them difficult to detect without specialized preprocessing strategies [3]. Although one-stage object detectors such as YOLO have demonstrated strong performance in real-time object detection tasks, their effectiveness in mammography remains limited due to the extremely small size of pathological findings relative to the full image resolution. Mammographic lesions are often characterized by subtle appearance, low contrast boundaries, and complex surrounding tissue structures, making accurate localization particularly challenging in dense breast images [4]. In addition, mammography images typically have ultra-high spatial resolution, whereas standard deep learning pipelines require aggressive image downsampling due to computational constraints. As a result, diagnostically important fine-grained features may become indistinguishable or disappear entirely during resizing, leading to reduced sensitivity for very small lesion detection [5]. The main contribution of this work lies in the application and evaluation of a practical patch-based YOLO detection pipeline for small-object detection in mammography. The study further provides a comparative analysis of model performance before and after tiling preprocessing. It examines the impact of patch-based training on the detection of malignant lesions in dense, high-resolution breast images.

This study aims to evaluate whether a patch-based preprocessing strategy can improve the detection performance of YOLO models for malignant lesion localization in mammography images.

II. RELATED WORK

AI has significantly transformed breast cancer diagnostics, providing robust tools for the automated analysis of mammograms, ultrasounds, and histopathology images. Within this domain, the You Only Look Once (YOLO) family of models has gained substantial traction due to its optimal balance of speed and accuracy. Mohammed and Ekmekci [6] introduced a YOLO-based multiscale parallel CNN to enhance the differentiation of malignant features. At the same time, Prinzi, Insalaco, Orlando, Gaglio, and Vitabile [7] demonstrated the efficacy of YOLO architectures specifically for real-time mammographic analysis. This versatility extends to other modalities; for instance, Ayuningtyas et al. [8] successfully implemented a YOLO-based framework for early detection in ultrasonography. Furthermore, Abdikenov, Rakishev, Orazayev, and Zhaksylyk [9] underscored that leveraging transfer learning is a critical strategy for enhancing detection performance, enabling models to adapt to the specific complexities of radiological data even with limited training data. Despite these advancements, a persistent bottleneck in medical imaging remains the "small object problem." As noted in the comprehensive reviews by Zhao, Zheng, Xu, and Wu [10] and Mirzaei, Nezamabadi-Pour, Raoof, and Derakhshani [11], standard object detectors often struggle when target lesions occupy a minute fraction (often less than 0.1%) of a high-resolution input. In such cases, global image down-sampling leads to a catastrophic loss of spatial information, effectively causing small anomalies to disappear. To mitigate this, patch-based approaches have been widely adopted to maintain local resolution. Hou et al. [12] and Ashour et al. [13] demonstrated that dividing large-scale images such as whole-slide tissue images or mammograms into smaller segments enables more effective feature representation and more manageable computational costs. The technical justification for this approach is often referred to as the "power of tiling." Unel, Özkalaycı, and Çığla [14]

provided empirical evidence that dividing high-resolution frames into overlapping patches significantly improves the performance of detectors such as YOLO by increasing the relative scale of small objects. This methodology is further supported by recent innovations such as Slicing Aided Hyper Inference (SAHI). Duwal and Dahal [15] highlighted how image slicing can overcome resolution disparities that typically cause standard models to fail. While some researchers, such as Manolakis, Bizopoulos, Lalas, and Votis [16], have proposed multi-stage frameworks that combine YOLO with segmentation networks to address these complexities, single-stage tiling pipelines remain highly attractive for their efficiency.

However, the transition from whole-image to patch-based learning introduces unique methodological hurdles. When an image is subdivided, the resulting dataset often suffers from extreme class imbalance, as most tiles may contain only healthy tissue. Bae and Ban [17] addressed this by proposing class-balanced random patch training to prevent model bias. Additionally, Ciga et al. [18] cautioned against the potential loss of global spatial context and feature inconsistency across tile boundaries. To ensure clinical viability, Shawi, Kilanava, and Sakr [19] emphasized the need for interpretability in these localized models, arguing that practitioners must understand the basis of patch-level predictions. Collectively, these studies illustrate that while YOLO combined with strategic tiling offers a potent solution for detecting small-scale anomalies, success depends on the careful optimization of tile size, overlap, and class distribution. There remains a critical need for streamlined pipelines that can achieve high recall for small malignant lesions while maintaining the inference speeds required for practical clinical decision support. The trend toward using YOLO-based architectures as foundational preprocessing and standardization tools is further exemplified by recent specialized medical applications. Zhanegiz, Mamyrbayev, Momynzhanova, and Zhumazhan [20] demonstrated that YOLO-based cropping can be more efficient than traditional K-means clustering for the automated standardization of fundus images, highlighting the model's robustness in preparing complex medical data for downstream analysis. Similarly, the development of high-performance architectures like HiFormer by Mamyrbayev, Zhanegizov, Momynzhanova, and Keylan [21] underscores the industry's shift toward models that prioritize both speed and accuracy in detecting subtle pathological features, such as those found in glaucoma. These advancements in ophthalmological imaging mirror the challenges faced in mammography, where precise localization and image standardization are paramount. By integrating a patch-based tiling strategy within the YOLOv8 framework, the present study aligns with these emerging trends, focusing on the critical need for preprocessing pipelines that enhance the visibility of small malignant lesions while maintaining the computational efficiency required for modern clinical environments. Recent studies on the INbreast dataset further demonstrate the growing interest in improving small-lesion detection in mammography using advanced deep learning architectures. Several approaches have explored transformer-based detectors, multi-scale feature pyramids, and hybrid segmentation–detection frameworks to improve sensitivity for subtle abnormalities. For example, Aly, Marey, El-Sayed, and Tolba [22] evaluated YOLO-based detection on the INbreast dataset and reported strong performance for mass detection in full-field digital mammograms. At the same time, Prinzi, Insalaco, Orlando, Gaglio, and Vitabile [23] investigated multiple YOLO architectures combined with transfer learning for mammographic lesion detection using both INbreast and CBIS-DDSM datasets. Similarly, recent multi-stage frameworks integrating YOLO with segmentation networks have shown promising results for mammographic mass localization and segmentation [24]. Nevertheless, many of these approaches rely on computationally intensive multi-stage pipelines or require extensive preprocessing and annotation refinement, limiting their applicability in real-time clinical workflows. Benchmarking studies indicate that while standard YOLO-based mammography models can achieve high detection accuracy for larger abnormalities, their performance on very small malignant lesions remains limited due to aggressive image downsampling and insufficient local feature representation. Additional studies on mammographic lesion analysis further demonstrated that resizing high-resolution mammograms to standard detector input dimensions may suppress diagnostically important fine-grained features [25]. Existing small object detection studies have attempted to address this limitation using feature pyramid networks, attention mechanisms, transformer-based architectures, and multi-resolution inference strategies [26], but these methods often increase architectural complexity and computational overhead. In contrast to many existing approaches, the present study focuses on a practical patch-based preprocessing pipeline integrated directly into a YOLOv8

detection framework. The proposed methodology emphasizes computational efficiency, simplified implementation, and improved visibility of small lesions in high-resolution mammograms while preserving the advantages of single-stage object detection. Furthermore, this work provides a comparative evaluation of model performance before and after tiling-based preprocessing, enabling a more direct assessment of how patch-based learning affects recall, precision, and localization performance in malignant lesion detection.

Beyond detection accuracy, the increasing integration of artificial intelligence into medical imaging pipelines also raises important concerns regarding data privacy, secure storage, and controlled sharing of sensitive patient information. Mammography datasets contain highly sensitive medical data, and the use of cloud-based AI systems or distributed training environments introduces additional security challenges, including unauthorized access, data leakage, and secure image transmission. Consequently, modern AI-enabled medical imaging frameworks must consider not only diagnostic performance but also mechanisms to preserve patient confidentiality and ensure secure medical image management. Recent security-oriented studies have proposed advanced privacy-preserving storage techniques, such as the Nested Secret Share (NSS) framework, for secure distributed image management. Galletta, Taheri, Celesti, Fazio, and Villari [26] demonstrated that Nested Secret Share approaches can improve the protection of sensitive image data while maintaining efficient accessibility in distributed environments. In subsequent work, the authors further addressed limitations of traditional Secret Share techniques by introducing enhanced nested architectures for secure data storage and sharing [27]. Such approaches may be highly relevant to future AI-assisted mammography systems, particularly in multi-center clinical collaborations, cloud-based diagnostic platforms, and federated medical AI infrastructures. Integrating secure storage and privacy-preserving mechanisms into patch-based mammography pipelines could improve the clinical feasibility, security, and trustworthiness of deep learning systems intended for real-world healthcare deployment.

III. METHODOLOGY

1. Dataset Description

This study used the publicly available INbreast dataset, hosted on Kaggle [28]. INbreast is a full-field digital mammography database widely used in breast cancer research. The dataset includes DICOM mammographic images, detailed lesion annotations, and clinical metadata.

The dataset provides several structured folders containing imaging data and annotations in various formats. An overview of the INBreast dataset components is presented in Table 1.

Table 1. Structure and contents of the INBreast dataset

File	Description
AllDICOMs	Original medical images in DICOM format
AllROI	Lesion annotations in OsiriX .roi format
AllXML	Lesion annotations in .xml format exported from OsiriX
MedicalReports	Radiologist narrative reports
PectoralMuscle	Manual annotations of the pectoral muscle boundary

File	Description
INbreast.xls/ INbreast.csv	Tabular metadata containing BI-RADS categories and clinical descriptors

In total, the dataset includes 412 DICOM mammograms, 343 ROI annotation files, 343 XML annotation files, 117 medical report files, and 201 pectoral muscle annotation files (both ROI and XML formats). In addition, a CSV file containing metadata for each examination is provided and was used to map images to diagnostic categories.

2. Data Preprocessing

The original mammograms in INBreast are provided in DICOM format. However, the YOLO object detection framework does not directly support DICOM images or OsiriX ROI/XML annotations. Therefore, all files were first downloaded to Google Drive, and a preprocessing pipeline was implemented in Google Colab.

As a first step, all DICOM images were converted to PNG format. The conversion was performed automatically in Colab to process the entire dataset at scale (hundreds of files), ensuring the complete preprocessing pipeline remained reproducible and integrated with subsequent model training.

Figure 1 shows two example PNG images demonstrating malignant breast lesions after DICOM conversion.

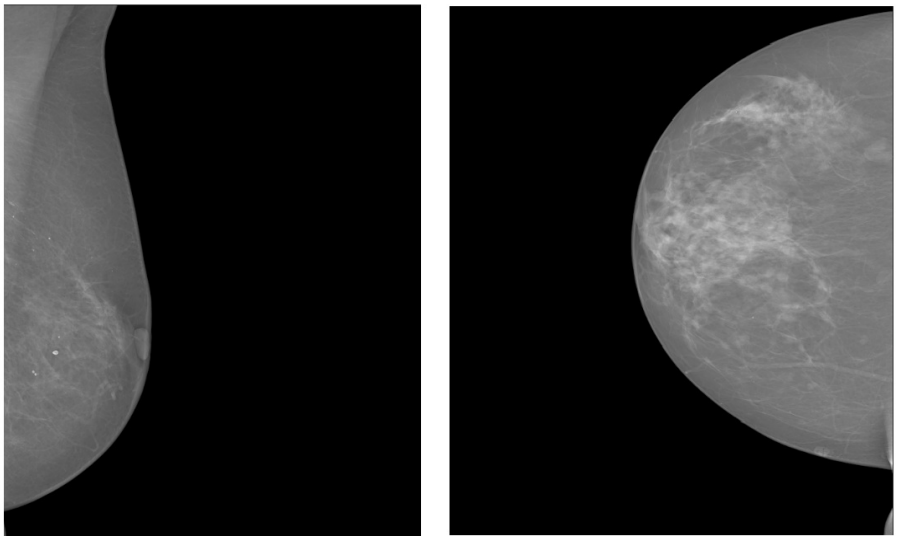


Figure 1. Examples of PNG images demonstrating malignant breast lesions after DICOM conversion

Because YOLO requires bounding-box annotations in TXT format, an automated annotation conversion pipeline was developed. INbreast provides lesion annotations in ROI/XML formats, which define polygon coordinates for each lesion, but do not explicitly include the mammographic projection (e.g., L/CC, L/MLO, R/CC, R/MLO). Therefore, each ROI/XML file was matched to the corresponding PNG image(s) using the patient identifier, which is present as the filename prefix in both the DICOM/PNG files and the annotation files. All PNG images whose filename started with the same patient ID as the XML file were considered valid matches.

For each XML annotation, the polygon points (x,y) were extracted, and the minimum and maximum coordinates were used to compute an axis-aligned bounding box. When a lesion contained a single point annotation, a small default bounding box (±10 px) was defined to preserve annotation consistency. Bounding boxes were then normalized to (x_center, y_center, width, height) based on the target PNG image resolution, following the YOLO format specification.

Lesion labels were derived from BI-RADS categories provided in the INbreast CSV metadata. Each metadata entry contains a "File Name" field that corresponds to the patient ID in the imaging files. BI-RADS categories were mapped into a binary classification schema: BI-RADS 1/2/3 were assigned to the benign class (label = 0), while BI-RADS 4, 4a, 5, and 6 were mapped to malignant (label = 1). This mapping follows clinical best practice, in which BI-RADS category 4 and above is considered "suspicious for malignancy."

A Python script running in Google Colab processed the entire dataset automatically. The script mounted Google Drive, read the CSV metadata (semicolon-delimited), mapped each patient ID to a binary class label, parsed XML files to extract polygon coordinates, computed bounding boxes, matched to PNG images, normalized coordinates, and exported YOLO-compatible TXT annotations. Multiple ROIs per image resulted in multiple annotation lines in the corresponding TXT file. The final label set contained 343 processed XML annotation files, resulting in a total of 7,316 YOLO bounding box labels, with no missing patient-image matches and no missing BI-RADS cases.

After the conversion, the final dataset consisted of 410 PNG images stored in the /INBreast_Release/PNG directory and 343 YOLO TXT annotation files stored in /INBreast_Release/labels. To verify annotation quality, a sample of images was visualized by rendering the generated bounding boxes directly onto the converted PNG mammograms. Figure 2 illustrates a representative visualization of the results after XML-to-YOLO conversion, confirming that the bounding boxes accurately localize the annotated lesions. A qualitative verification of the generated YOLO annotations is presented in Figure 2.

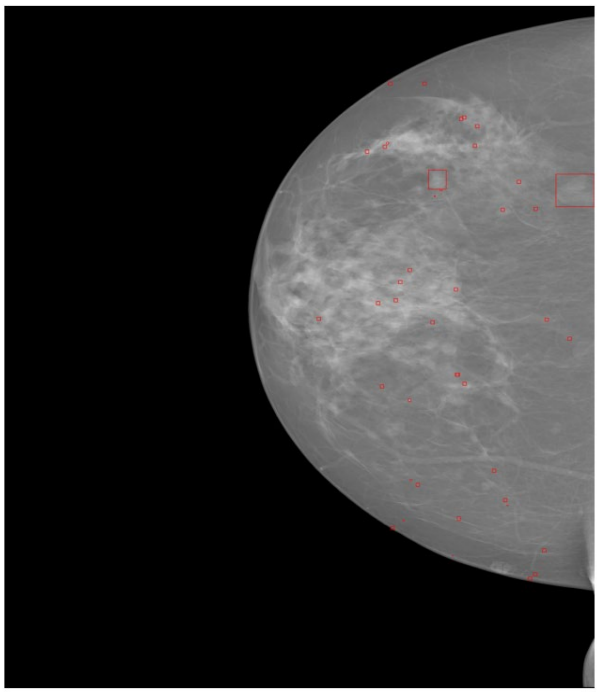


Figure 2. A representative example of the resulting visualization after XML-to-YOLO conversion

The dataset was partitioned into three subsets to support model development, hyperparameter tuning, and unbiased performance evaluation. A stratified split was applied at the image level with the following proportions: 70% for training, 20% for validation, and 10% for testing. After splitting, the final subset sizes were 287 images for training, 82 images for validation, and 41 images for testing. All subsets were mutually exclusive, and no patients were allowed to overlap across partitions.

3. Model Architecture

A YOLOv8 object detection model was selected for its optimized computational efficiency and high performance on small- and medium-scale datasets. The Ultralytics YOLOv8 implementation was installed and used. The training environment was configured by mounting Google Drive, verifying GPU availability, updating Ultralytics, and validating the dataset structure (images, labels, and data.yaml configuration file).

Model training was conducted using a single GPU in Google Colab, with an input resolution of 640×640 pixels for both training and validation images. The model was trained for 50 epochs. Batch size and image size parameters were adjusted dynamically if out-of-memory (OOM) errors occurred during initialization. Throughout training, the framework logged epoch-wise performance metrics including GPU memory usage, box loss, classification loss, and distribution focal loss. Upon training completion, the best-performing checkpoint (best.pt) was automatically selected based on the highest validation performance.

Following training, the model was evaluated on the held-out test subset (N=41 images). Example visualizations were generated to assess detection performance qualitatively, and the trained model was exported for downstream inference and analysis.

4. Patch-based training

Mammography images are typically very large (≈3000×4000–4000×6000 px), while cancer lesions can be extremely small (20–50 px), low-contrast, and visually similar to background tissue, often occupying less than 0.1% of the image area. Standard object detectors such as YOLO and other CNN-based models struggle to detect such small objects, especially after downsampling the image to 640×640 or 1024×1024, where the lesion may shrink to a few pixels or disappear entirely. In our dataset, the average bounding box area was 0.0004 (≈0.04% of the image).

Tiling (patch-based training) addresses these challenges as follows:

1. Image splitting: Large images are divided into smaller overlapping tiles (e.g., 640×640 with 20–30% overlap). A typical 4000×3000 image generates ~30–50 tiles, effectively increasing the dataset size and standardizing input dimensions.
2. Bounding box recalculation: Only boxes intersecting a tile are included; coordinates are adjusted relative to the tile.
3. Lesion magnification: Small lesions (e.g., 30–40 px) are enlarged on tiles (≈200–300 px), making them detectable by YOLO.
4. Foreground-background balance: Tiling reduces extreme background dominance, allowing tiles to contain tissue patterns or pathologies, improving model learning of relevant features.
5. Dataset expansion: For example, after tiling, the train/validation/test sets increased from 287/82/41 images to 1657/487/270 tiles, resulting in 12,057 tiles with objects (class distribution: 8462 benign, 3595 malignant), a 20–40× increase in effective training samples.

IV. RESULTS

The trained YOLOv8 model (best.pt) was loaded and used for inference on the held-out test set. The model object provided multiple functionalities. Predictions were performed on all test images, and the results were saved as annotated images with bounding boxes overlaid and as text files containing the coordinates of predicted objects. Annotated images enable visual inspection of lesion-detection accuracy, while TXT files provide structured data for quantitative evaluation. Figure 3 illustrates predicted bounding boxes on test images, demonstrating the model's ability to localize breast lesions accurately.

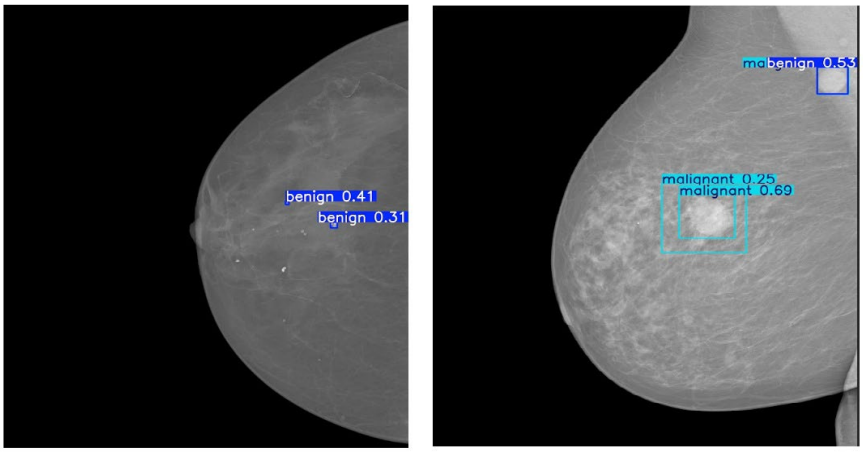


Figure 3. Examples of the predicted bounding boxes on test images

The trained YOLOv8 model was evaluated on the held-out test set using the standard validation procedure implemented in the Ultralytics framework. The evaluation results are summarized in Table 2.

Table 2. Detection performance and inference efficiency of YOLOv8 on the INBreast dataset

Metric	Precision	Recall	F1-score
Value	30%	3-4%	5-7%

The obtained results indicate limited detection performance, particularly for small malignant lesions in high-resolution mammographic images. To identify the underlying causes of low precision, recall, and mAP scores, an error analysis was conducted. Based on its findings, several improvements were introduced, including tiling-based preprocessing, enhanced data augmentation, and adjustments to the training pipeline to mitigate class imbalance and improve small object detection. The original dataset contained 410 mammograms and 343 YOLO-format annotation files.

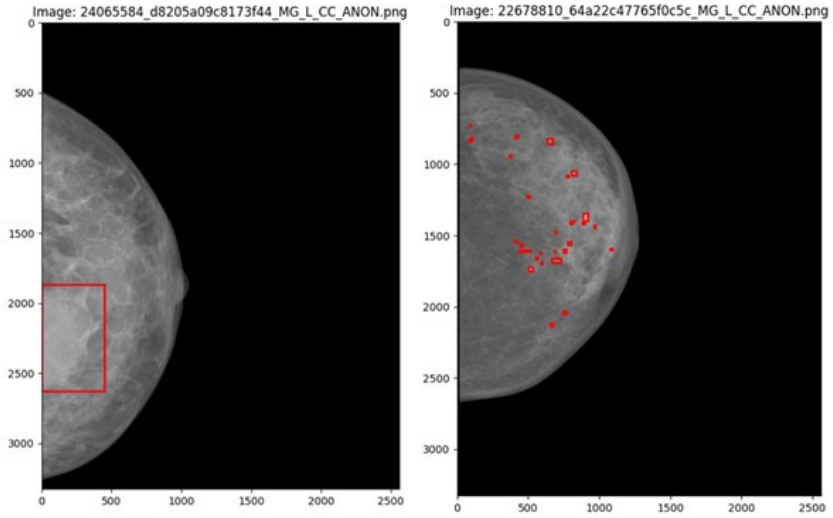


Fig. 1. Visualization of Random Bounding Boxes on Mammographic Images

Data validation revealed that 67 images lacked corresponding annotations; these were removed to ensure consistency, resulting in a final dataset of 343 image–label pairs. To further characterize the dataset, a visual inspection of regions of interest and an analysis of bounding box size distribution were performed. This confirmed the accuracy of the annotations and provided insights into lesion-scale variability. The results indicated a high prevalence of small objects, supporting the need for patch-based processing strategies such as tiling to improve detection performance, as illustrated in Fig. 4. Additionally, the distribution analysis provided insights into the relative sizes of objects in the images, helping assess whether small-object detection might be an issue and whether tiling or patch-based training would improve model performance. The visual and statistical examination of bounding boxes thus informed subsequent model design and training strategies, as illustrated in Fig. 4, which shows randomly selected bounding boxes on mammographic images. The bounding box size distribution analysis indicated that lesions occupy an extremely small proportion of the original mammographic images. The normalized average bounding box area was 0.0004, with minimum and maximum values of 0.0000 and 0.0896, respectively. These results confirm that the dataset is dominated by small objects, with only a limited number of relatively large lesions. The presence of near-zero bounding box areas suggests potential labeling noise or annotation inconsistencies. At the same time, the overall distribution highlights the inherent difficulty of detecting small-scale structures in high-resolution images. These observations support the use of patch-based or tiling strategies to improve lesion visibility and model sensitivity. After applying the tiling procedure, the dataset was expanded to 12,057 annotated image tiles, comprising 8,462 benign and 3,595 malignant instances. The normalized bounding box area within tiles had a mean of 0.0108, with minimum and maximum values of 0.0000 and 1.0000, respectively, indicating a substantial increase in the relative lesion scale after preprocessing. This transformation improves the object-to-background ratio and ensures better utilization of detection features within the YOLO architecture. As a result, small lesions are more effectively preserved within feature pyramids and anchor-based representations, enhancing the model's ability to learn fine-grained pathological patterns. Representative examples of tiled images with bounding box annotations are shown in Fig. 5.

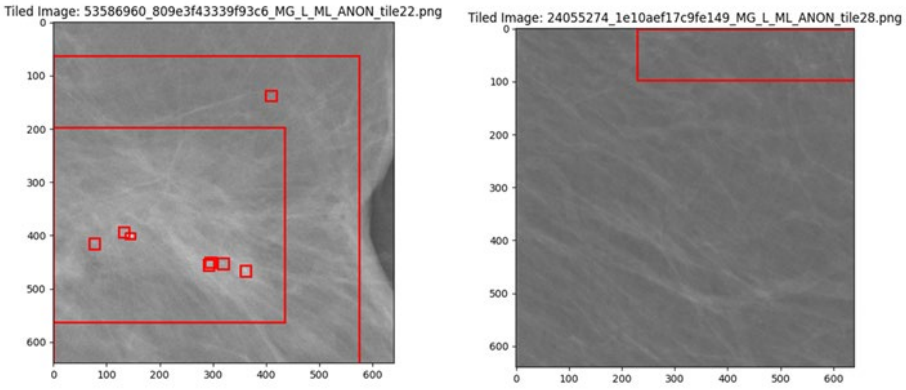


Fig. 2. Qualitative inspection of bounding box annotations on multiple image tiles

The YOLO model was trained and validated on the tiled mammography dataset. After training, the model achieved the following overall metrics: Precision = 0.277, Recall = 0.450, mAP50 = 0.351, and mAP50–95 = 0.212, with a low training loss of 0.0004, indicating effective convergence. The trained model and results were saved. Inference speed was measured at 2.18 s per image, with preprocessing and postprocessing times of 0.24 s and 3.78 s, respectively. Validation was performed on 487 images containing 2,121 lesion instances. Overall detection metrics on the validation set were: Precision = 0.281, Recall = 0.454, mAP50 = 0.352, and mAP50–95 = 0.213. For class-specific performance, benign lesions achieved higher detection scores (Precision = 0.336, Recall = 0.745, mAP50 = 0.574), while malignant lesions were more challenging to detect (Precision = 0.226, Recall = 0.164, mAP50 = 0.130). The validation results confirm that tiling and patch-based training improved the model's ability to detect small lesions, particularly benign ones, although detecting malignant lesions remains more difficult. To qualitatively assess the model predictions, several test images were visualized with predicted bounding boxes overlaid on the original mammograms. Fig. 6 shows examples of correctly detected malignant and benign lesions, as well as some false positives.

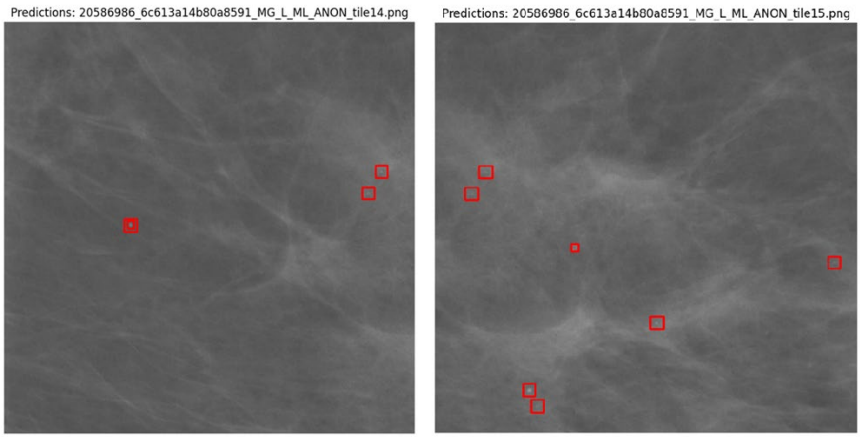


Fig. 3. Examples of YOLOv8 predictions on test images

The initial dataset contained 410 mammography images with 343 YOLO-format annotation files. A preliminary inspection revealed 67 unlabeled images, which were excluded to ensure data consistency. The cleaned dataset thus consisted of 343 images with complete annotations. A detailed analysis of bounding box (bbox) sizes revealed that lesions were predominantly small relative to the full image size. The average bbox area was 0.0004 (≈0.04% of the original image), with minimum and maximum areas of 0.0000 and 0.0896, respectively. Near-zero-size boxes indicated potential labeling artifacts, while the rare larger lesions represented less than 10% of the image area. Visual inspection of ROIs confirmed that the majority of bounding boxes accurately corresponded to malignant lesions. Given the relatively small size of the lesions, tiling (patch-based training) was applied. Large images (≈4000×3000 px) were split into 640×640 tiles with 20–30% overlap, resulting in ~30–50 tiles per image. This approach increased the effective dataset size and enlarged the relative size of lesions within each tile, making them more detectable by YOLO. After tiling, the dataset contained 12,057 tiles with annotated objects, distributed as 8,462 benign and 3,595 malignant lesions. Average bbox area per tile increased to 0.0108, with a minimum of 0.0000 and a maximum of 1.0000, reflecting the enlarged object representation on tiles. The tiling procedure also reduced the extreme background-to-object imbalance inherent in mammography images.

Table 3 presents the detection performance of the YOLOv8 model evaluated on the training, validation, and test sets.

Table 3. Detection metrics

Metric	Precision	Recall	F1-score
Train set	0,277	0,450	0,343
Validation set	0,281	0,454	0,347
Test set	0.477	0.622	0.540

The model achieved a recall of 62.2%, indicating that most true lesions were detected, while the precision of 47.7% reflects false-positive detections. The overall F1-score of 0.54 represents a balanced performance given the challenging nature of the mammographic images, which often contain small lesions and artifacts. These results highlight the effectiveness of patch-based tiling and YOLOv8 for breast lesion detection and suggest potential improvements through larger datasets, class balancing, and advanced data augmentation techniques.

V. CONCLUSION

This study presented a YOLO-based framework for malignant lesion detection in high-resolution mammography images, enhanced with a patch-based tiling preprocessing strategy. The proposed approach was designed to mitigate key challenges associated with mammographic analysis, including extreme class imbalance, low lesion contrast, and the disproportionately small size of pathological structures relative to full-image resolution. The incorporation of tiling significantly improved the effective representation of small lesions by increasing their relative scale and improving local feature visibility during training and inference. Experimental results demonstrate a substantial increase in detection sensitivity, with a marked improvement in recall over the baseline full-image configuration, indicating enhanced capability to identify small malignant regions. Despite these improvements, the model still exhibits limitations in precision and false-positive generation, primarily due to background tissue complexity and class imbalance. These findings indicate that while patch-based preprocessing improves sensitivity, further optimization is required to balance detection robustness and specificity. Overall, the results confirm that tiling is a critical preprocessing strategy for improving small-object detection in

mammography using YOLO-based architectures. Future work will focus on extending the experimental evaluation to additional public mammography datasets to assess generalizability, as well as conducting comprehensive comparisons with state-of-the-art detection architectures. Further research will also investigate imbalance-aware learning strategies, advanced post-processing techniques, and hybrid modeling approaches that integrate local patch-based inference with global contextual representations to improve both robustness and diagnostic reliability in clinical applications.

ACKNOWLEDGEMENT

The work was supported by grant funding from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, BR24993001: Creation of a large language model (LLM) to maintain the implementation of the Kazakh language and increase technological progress.

REFERENCES

[1]. World Health Organization, "Breast cancer." World Health Organization Breast Cancer Fact Sheet, last accessed Apr. 15, 2026.

[2]. Y. Khourdifi, A. El Alami, M. Zaydi, Y. Maleh, and O. Er-Remyly, "Early breast cancer detection based on deep learning: An ensemble approach applied to mammograms," BioMedInformatics, vol. 4, no. 4, pp. 2338–2373, 2024.

[3]. G. H. Aly, M. Marey, S. A. El-Sayed, and M. F. Tolba, "YOLO-based breast masses detection and classification in full-field digital mammograms," Comput. Methods Programs Biomed., vol. 200, p. 105823, 2021.

[4]. H. Abudukelimu, Y. Gao, A. Abulizi, M. Musideke, S. Wu, M. Wang, et al., "DVF-YOLO-Seg: A two-stage breast mass segmentation model with enhanced feature extraction and small lesion detection," Digital Health, vol. 11, pp. 1–18, 2025.

[5]. K. Rangarajan, A. Gupta, S. Dasgupta, U. Marri, A. K. Gupta, S. Hari, et al., "Ultra-high resolution, multi-scale, context-aware approach for detection of small cancers on mammography," Scientific Reports, vol. 12, no. 1, Art. no. 11622, 2022.

[6]. A. D. Mohammed and D. Ekmekci, "Breast cancer diagnosis using YOLO-based multiscale parallel CNN and flattened threshold swish," Appl. Sci., vol. 14, no. 7, p. 2680, 2024.

[7]. F. Prinzi, M. Insalaco, A. Orlando, S. Gaglio, and S. Vitabile, "A YOLO-based model for breast cancer detection in mammograms," Cogn—Comput., vol. 16, no. 1, pp. 107–120, 2024.

[8]. A. Ayuningtyas, H. Wintolo, A. D. W. Sumari, E. Setyaningsih, A. Pujiastuti, A. S. Honggowibowo, and A. Kusumaningrum, "The CNN model with YOLO architecture for ultrasonography images in early breast cancer detection," J. Appl. Data Sci., vol. 6, no. 2, pp. 1116–1128, 2025.

[9]. B. Abdikenov, D. Rakishev, Y. Orazayev, and T. Zhaksylyk, "Enhancing breast lesion detection in mammograms via transfer learning," J. Imaging, vol. 11, no. 9, p. 314, 2025.

[10]. Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 11, pp. 3212–3232, 2019.

[11]. B. Mirzaei, H. Nezamabadi-Pour, A. Raoof, and R. Derakhshani, "Small object detection and tracking: A comprehensive review," Sensors, vol. 23, no. 15, p. 6887, 2023.

[12]. L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2424–2433.

[13]. L. S. Ashour, A. A. Mohammed, M. M. A. Zaid, P. Sumari, A. K. J. Al-Nussairi, S. A. Al-Shammari, and S. T. Abdulmunem, "Non-overlapping patch-based pre-trained CNN for breast cancer classification," Iraqi J. Comput. Sci. Math., vol. 6, no. 2, p. 29, 2025.

[14]. F. O. Ünél, B. O. Özkalaycı, and C. Çığla, "The power of tiling for small object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2019, pp. 0–0.
[15]. R. Duwal and A. Dahal, "Improved FMCG quality inspection through small object detection using YOLOv5 and SAHI-based image slicing," Authorea Preprints, 2025.
[16]. D. Manolakis, P. Bizopoulos, A. Lalas, and K. Votis, "A two-stage lightweight deep learning framework for mass detection and segmentation in mammograms using YOLOv5 and Depthwise SegNet," J. Imaging Inform. Med., pp. 1–16, 2025.
[17]. Y. Bae and Y. Ban, "Class-balanced random patch training to address class imbalance in tiling-based farmland classification," Appl. Sci., vol. 15, no. 13, p. 7056, 2025.
[18]. O. Ciga, T. Xu, S. Nofech-Mozes, S. Noy, F. I. Lu, and A. L. Martel, "Overcoming the limitations of patch-based learning to detect cancer in whole slide images," Sci. Rep., vol. 11, no. 1, p. 8894, 2021.
[19]. R. E. Shawi, K. Kilanava, and S. Sakr, "An interpretable semi-supervised framework for patch-based classification of breast cancer," Sci. Rep., vol. 12, no. 1, p. 16734, 2022.
[20]. A. Zhanegiz, O. Mamyrbayev, K. Momynzhanova, and N. Zhumazhan, "Effective clustering with K-means or efficient cropping with YOLO as a preprocessing step for automated fundus image standardization," in Proc. Sixth Int. Conf. Comput. Vis. Inf. Technol. (CVIT 2025), Proc. SPIE, vol. 13796, pp. 27–34, 2025.
[21]. O. Mamyrbayev, A. Zhanegizov, K. Momynzhanova, and A. Keylan, ″HiFormer: A fast and accurate deep learning model for segmentation and detection of glaucoma," J. Supercomput., vol. 82, no. 1, p. 13, 2025.
[22]. G. H. Aly, M. Marey, S. A. El-Sayed, and M. F. Tolba, "YOLO based breast masses detection and classification in full-field digital mammograms," Computer Methods and Programs in Biomedicine, vol. 200, Art. no. 105823, 2021.
[23]. F. Prinzi, M. Insalaco, A. Orlando, S. Gaglio, and S. Vitabile, "A YOLO-based model for breast cancer detection in mammograms," Cognitive Computation, vol. 16, no. 1, pp. 107–120, 2024.
[24]. S. J. Frank, "A deep learning architecture with an object-detection algorithm and a convolutional neural network for breast mass detection and visualization," Healthcare Analytics, vol. 3, Art. no. 100186, 2023.
[25]. S. A. Kamran, K. F. Hossain, A. Tavakkoli, G. Bebis, and S. Baker, "Swin-SFTNet: Spatial feature expansion and aggregation using Swin transformer for whole breast micro-mass segmentation," in Proc. IEEE 20th Int. Symp. Biomed. Imaging (ISBI), 2023, pp. 1–5.
[26]. A. Galletta, J. Taheri, A. Celesti, M. Fazio, and M. Villari, "Investigating the applicability of nested secret share for drone fleet photo storage," IEEE Transactions on Mobile Computing, vol. 23, no. 4, pp. 2671–2683, 2024.
[27]. A. Galletta, J. Taheri, M. Fazio, A. Celesti, and M. Villari, "Overcoming security limitations of secret share techniques: The nested secret share," in Proc. IEEE 20th Int. Conf. Trust, Security and Privacy in Computing and Communications (TrustCom), 2021, pp. 289–296.
[28]. INbreast Dataset, Kaggle. INbreast Dataset on Kaggle, last accessed Apr. 15, 2026.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

From Regulatory Compliance to Digital Trust: Integrating NIS2 and ISO/IEC 27701 In Transitional Cybersecurity Governance - The Case of Bosnia and Herzegovina

Edin Alić
HIFA-OIL d.o.o. Tešanj
Tešanj, Bosnia and Herzegovina
edin.alic@hifaoil.ba

Muhamed Ćosić
Centar za održivost i digitalizaciju
Centar EkoDigitus, Travnik
Bosnia and Herzegovina
drmuhamedcosic@gmail.com

Venan Hadžiselimović
HIFA-OIL d.o.o. Tešanj
Tešanj, Bosnia and Herzegovina
venan.hadziselimovic@hifaoil.ba

Abstract
Cybersecurity and privacy governance are now central to digital resilience, particularly in environments affected by growing cyber threats, regulatory fragmentation, and dependence on interconnected infrastructures. Although the NIS2 Directive and ISO/IEC standards pursue complementary objectives, their integrated operational implementation within transitional regulatory systems remains insufficiently explored. This paper develops and operationalizes a Digital Trust Governance Model integrating NIS2, ISO/IEC 27001, and ISO/IEC 27701 into a unified governance maturity framework. Using comparative legal analysis and governance gap analysis, the study evaluates the cybersecurity and privacy governance structure of Bosnia and Herzegovina with emphasis on risk management, incident reporting, CSIRT coordination, supervisory capacity, and management accountability. The proposed model conceptualizes digital trust as a measurable governance outcome derived from four interdependent dimensions: normative precision, institutional capacity, sanctioning capability, and technical and organizational implementation. To improve analytical applicability, the framework introduces qualitative maturity indicators for assessing operational cyber resilience and governance effectiveness. The findings reveal significant asymmetry between privacy governance maturity and cybersecurity governance capability, particularly regarding institutional coordination, enforcement consistency, and implementation capacity. The paper contributes to the literature by integrating regulatory, institutional, enforcement, and operational dimensions into a single governance maturity framework.

Keywords: Digital Trust Governance; Cybersecurity Governance; NIS2; ISO/IEC 27701; ISO/IEC 27001; Governance Maturity; Cyber Resilience; Privacy Governance; Regulatory Compliance; Bosnia and Herzegovina

I. INTRODUCTION

Cybersecurity and personal data protection are core components of digital governance because interconnected systems and critical infrastructures expose organizations to legal, operational, economic and reputational consequences when incidents occur. Cybersecurity has therefore evolved from a narrow IT function into a strategic governance responsibility requiring coordinated regulation, institutional capacity, executive accountability and effective implementation of controls. The EU addresses these challenges through GDPR [1] and NIS2 [2]. GDPR strengthens privacy, lawful processing and accountability, while NIS2 broadens cybersecurity duties concerning risk management, incident reporting, supply-chain security, business continuity, supervision and management responsibility. ISO/IEC 27001 and ISO/IEC 27701 translate these obligations into operational management systems. Their integration is especially relevant in transitional systems such as Bosnia and Herzegovina, where privacy governance is comparatively more consolidated while cybersecurity coordination, incident reporting, CSIRT integration and management accountability remain fragmented [3], [4]. This paper therefore evaluates the alignment of NIS2, ISO/IEC 27001 and ISO/IEC 27701 with the legal and

institutional framework of Bosnia and Herzegovina through comparative and gap analysis, focusing on risk management, incident reporting, CSIRT coordination, supervision and executive accountability. The study treats digital trust as a measurable governance outcome of normative precision, institutional capacity, sanctioning capability and technical-organizational implementation, and proposes a scalable model for identifying structural weaknesses in transitional regulatory systems.

II. RESEARCH CONTRIBUTIONS

The study contributes an integrated Digital Trust Governance Model linking NIS2 with ISO/IEC 27001 and ISO/IEC 27701, a qualitative maturity assessment framework, and a comparative application to Bosnia and Herzegovina.

III. LITERATURE REVIEW

1. Cybersecurity Governance and Regulatory Transformation

Cybersecurity governance increasingly emphasizes resilience, executive responsibility and institutional coordination rather than isolated technical controls. NIS2 reflects this shift through mandatory risk management, incident reporting, supply-chain security, business continuity and supervisory accountability, while NIST CSF 2.0 supports continuous governance through identify, protect, detect, respond and recover functions [5].

2. Privacy Governance and Data Protection Frameworks

Privacy governance similarly emphasizes accountability, transparency and demonstrable safeguards. GDPR established a comprehensive framework, but privacy and cybersecurity are still often governed separately. In Bosnia and Herzegovina this separation can weaken coordination when incidents involve both infrastructure disruption and personal data compromise [6].

3. ISO Standards and Operational Implementation

ISO/IEC 27001 provides a risk-based ISMS for identifying risks, implementing controls and continually improving information security. It operationalizes many NIS2 requirements, including incident management, continuity, supplier security and access control. ISO/IEC 27701:2025 extends this structure with a Privacy Information Management System (PIMS), connecting privacy obligations with cybersecurity governance and organizational accountability.

4. Digital Trust and Governance Effectiveness

Digital trust denotes confidence in the security, reliability, accountability and integrity of digital systems. Governance-oriented approaches increasingly view it as dependent on legal certainty, institutional effectiveness, implementation capability and operational resilience [7]. In fragmented or transitional systems, trust is strongly affected by coordination, supervisory consistency and response capacity; this study therefore links digital trust to normative precision, institutional capacity, sanctioning capability and technical implementation [8-11].

5. Governance Maturity and Cyber Resilience Assessment

Recent governance and resilience literature moves beyond formal compliance toward measurable prevention, detection, response and recovery capability. NIS2, DORA and ENISA guidance reinforce integrated governance, continuous risk management, supervisory capacity and incident coordination [12-15]. Despite these developments, few approaches integrate cybersecurity governance, privacy governance, enforcement capability and operational implementation into one maturity framework for transitional regulatory systems.

6. Research Gap

Existing literature addresses cybersecurity regulation, privacy governance and international standards extensively, but important integration gaps remain. Cybersecurity and privacy are often analyzed separately, alignment between NIS2 and ISO/IEC 27001/27701 is insufficiently examined, and transitional regulatory systems remain underrepresented. Digital trust is also rarely operationalized as a governance outcome combining legal, institutional, enforcement and technical dimensions.This study addresses these gaps through an integrated Digital Trust Governance Model and a qualitative maturity assessment applied to Bosnia and Herzegovina.

IV. METHODOLOGY

1. Research Design

The study uses an interdisciplinary qualitative design combining legal analysis, cybersecurity governance theory and operational compliance assessment. It evaluates not only formal rules but also institutional effectiveness and implementation capability in Bosnia and Herzegovina.

2. Research Approach

Three complementary methods are used: normative-dogmatic analysis, comparative analysis and governance gap analysis.

3. Normative Analysis

Normative analysis examines cybersecurity, privacy, risk-management, incident-reporting and supervisory obligations in European and national frameworks.

4. Comparative Analysis

Comparative analysis evaluates functional alignment between NIS2 and Bosnia and Herzegovina in cyber risk management, incident reporting, CSIRT coordination, governance accountability and supervisory capacity.

5. Gap Analysis

Gap analysis compares regulatory requirements with existing implementation mechanisms to identify discrepancies between formal alignment and practical governance capability.

6. Analytical Framework

The analytical matrix compares NIS2 obligations, ISO/IEC 27001 implementation mechanisms, ISO/IEC 27701 privacy requirements, the BiH legal-institutional framework and the level of operational alignment.

Alignment is classified as fully aligned, partially aligned or not aligned, enabling structured assessment of governance maturity across cybersecurity and privacy domains.

7. Data Sources

Sources include EU regulatory instruments, international standards, national legislation, institutional reports, cybersecurity-governance literature and public policy documents.

Primary sources are NIS2, GDPR, ISO/IEC 27001 and ISO/IEC 27701; secondary sources cover cybersecurity governance, regulatory effectiveness, risk governance and digital trust.

8. Evaluation Criteria

Assessment criteria include normative clarity, institutional coordination, operational feasibility, incident-reporting capability, supervisory consistency, implementation maturity and accountability.

9. Conceptual Governance Model

The conceptual model evaluates four interdependent dimensions: normative precision (N), institutional capacity (I), sanctioning capability (S) and technical implementation (T).

Governance effectiveness is treated as their interaction; weakness in any single dimension can materially reduce overall maturity.

10. Governance Maturity Scoring Framework

Each model dimension is assessed on a five-level qualitative maturity scale from fragmented to integrated governance, reflecting development, operational integration and implementation effectiveness.

Table 1. Governance Maturity Scale

Score	Level	Description
1	Fragmented	Governance mechanisms are largely undeveloped, inconsistent, or absent
2	Reactive	Basic governance mechanisms exist but remain limited and primarily reactive
3	Developing	Governance structures are partially implemented with moderate operational capability
4	Managed	Governance mechanisms are formally integrated and operationally coordinated

5	Integrated	Governance structures demonstrate high maturity, coordination, enforcement, and resilience capability

Source: Author's qualitative assessment framework based on cybersecurity governance maturity literature.

The scoring is qualitative and comparative rather than statistically deterministic; it structures governance assessment rather than claiming precise measurement.

11. Limitations of the Study

The study is limited by BiH's fragmented competences, reliance on public legal and institutional sources, and its qualitative focus rather than direct measurement of incidents or organizational performance.

V. REGULATORY AND STANDARD FRAMEWORK

1. NIS2 and Cybersecurity Governance

NIS2 establishes a governance-oriented EU cybersecurity framework based on risk management, resilience, supervision, institutional coordination and management accountability [16-18]. Covered entities must implement appropriate technical, operational and organizational measures, including incident response, business continuity, supply-chain security, vulnerability management, security monitoring and executive accountability. NIS2 also strengthens time-bound incident reporting, national CSIRT coordination and cross-border cooperation, while explicitly placing cyber-risk responsibility on management bodies.

2. Privacy Governance and GDPR-Based Frameworks

GDPR is the central EU privacy framework, emphasizing lawful processing, transparency, purpose limitation, integrity, confidentiality, accountability, breach notification and technical-organizational safeguards. Its accountability model requires organizations to demonstrate effective safeguards. In Bosnia and Herzegovina, privacy governance is relatively more consistent than cyber governance, but remains insufficiently integrated with broader resilience and incident-response structures.

3. ISO/IEC 27001 and Information Security Governance

ISO/IEC 27001 provides a risk-based ISMS for establishing, maintaining and continually improving information security through risk assessment, incident management, access control, asset and supplier protection and business continuity. It therefore serves as an operational mechanism for implementing many NIS2 cybersecurity obligations.

4. ISO/IEC 27701 and Privacy Information Management

ISO/IEC 27701:2025 extends the ISO/IEC 27001 management-system structure with PIMS requirements for PII governance, controller/processor obligations, privacy risk, data-subject rights and accountability. This creates an operational bridge between security and privacy governance and supports more consistent implementation where institutional responsibilities remain separated.

5. Integrated Governance Perspective

NIS2 and ISO standards are complementary: NIS2 defines regulatory expectations, supervision, reporting and coordination, while ISO/IEC 27001 and ISO/IEC 27701 provide operational security and privacy controls. Their integration supports a transition from formal compliance to operational governance maturity and measurable digital trust.

6. Structural Limitations in Transitional Regulatory Systems

Transitional systems often show a gap between legal reform and implementation capability.

In Bosnia and Herzegovina the main limitations are fragmented competences, decentralized coordination, inconsistent supervision, limited incident coordination and uneven implementation maturity. These conditions create an implementation gap that can only be reduced through stronger coordination, operational resilience and integrated governance [19-21].

7. Gap Analysis

8. Methodological Basis of the Gap Analysis

The gap analysis compares NIS2, ISO/IEC 27001 and ISO/IEC 27701 with the BiH legal and institutional framework, focusing on functional capability rather than formal equivalence.

9. Scope of the Analysis

Five domains are assessed: cyber risk management, incident reporting, CSIRT coordination, supervision and enforcement, and management accountability.

10. Comparative Mapping

Table 2. Gap Analysis Matrix: Nis2, Iso Standards And Bih Framework

Requirement	**NIS2**	**ISO/IEC 27001**	**ISO/IEC 27701**	**BiH Framework**	**Gap Level**
Risk management	Mandatory risk-based approach	ISMS risk controls	Privacy risk controls	Partially regulated	Partial
Incident reporting	Time-bound reporting	Incident management controls	Data breach procedures	Not unified	High
CSIRT coordination	National coordination required	Indirect support	Indirect support	Fragmented	High
Supervision and enforcement	Strong oversight	Not directly applicable	Not directly applicable	Limited and uneven	Partial

Management accountability	Explicit executive responsibility	Governance responsibility	Privacy roles	Weakly defined	High

Source: Author's qualitative assessment framework based on cybersecurity governance maturity literature.

The matrix shows that the principal weaknesses concern institutional coordination and operational implementation rather than the mere existence of legal norms.

11. Key Findings

The analysis reveals a clear asymmetry between comparatively stronger privacy governance and more fragmented cybersecurity governance in Bosnia and Herzegovina. Risk management is only partially aligned with NIS2, and a fully integrated national cyber-risk framework is absent. Incident reporting remains a major gap because privacy-breach mechanisms are clearer than unified reporting for broader cybersecurity incidents. CSIRT capacities exist but are not sufficiently integrated into a consolidated cross-sector coordination and response system. Supervision and enforcement are uneven, with no consistently centralized cybersecurity authority combining clear powers and operational capacity. Management accountability is also weakly institutionalized compared with NIS2's explicit board-level responsibility.

12. Structural Causes of Identified Gaps

The gaps are structural as well as normative, reflecting fragmented competences, limited resources, weak supervisory capacity and insufficient integration of privacy and cyber governance.

13. Separated cyber and privacy governance creates incomplete risk assessment, inconsistent incident response and weaker inter-agency coordination.

An integrated governance approach should combine NIS2 obligations, ISO/IEC 27001 security controls, ISO/IEC 27701 privacy controls, institutional coordination and management accountability.

Such a framework should combine:

- NIS2 regulatory obligations,
- ISO/IEC 27001 security controls,
- ISO/IEC 27701 privacy controls,
- institutional coordination mechanisms,

The gap analysis shows that formal alignment does not by itself create resilience. The proposed model therefore connects regulatory obligations, institutional structures, enforcement and implementation as mutually dependent components of digital trust. This transition would enable Bosnia and Herzegovina to move from formal regulatory alignment toward measurable implementation capacity and digital trust maturity. The identified gaps directly justify the four dimensions of the proposed model: weak incident reporting reflects insufficient normative precision, fragmented CSIRT coordination indicates limited institutional capacity, uneven supervision points to weak sanctioning capability, while inconsistent organizational controls reveal gaps in technical implementation.

VI. INTEGRATED DIGITAL TRUST GOVERNANCE MODEL

1. Rationale for Model Development

The gap analysis demonstrates that formal regulatory alignment does not necessarily produce effective cyber resilience or privacy protection. In transitional regulatory systems, legal norms may exist without sufficient institutional capacity, enforcement consistency, or operational implementation. This creates a gap between compliance as a formal requirement and governance as a practical capability. For this reason, the study proposes an integrated Digital Trust Governance Model designed to connect regulatory obligations, institutional structures, enforcement mechanisms, and technical implementation into a unified analytical framework. The model is based on the assumption that digital trust cannot be produced by regulation alone. It emerges from the interaction between legal clarity, institutional effectiveness, supervisory credibility, and operational security maturity.

2. Normative precision concerns clear and predictable duties for risk management, incident reporting, data protection, management responsibility and supervision.

The proposed model consists of four interdependent variables:

3. Normative Precision (N)

Institutional capacity concerns the human, technical, financial and coordination capability of supervisory and response bodies.

4. Institutional Capacity (I)

Institutional capacity refers to the ability of supervisory and response bodies to implement, monitor, and enforce regulatory obligations. It includes human resources, technical expertise, budgetary stability, operational independence, and inter-agency coordination. Without sufficient institutional capacity, even well-designed regulations may remain ineffective in practice.

5. Sanctioning capability concerns credible, proportionate and dissuasive enforcement, including penalties, transparency and corrective action.

Sanctioning capability refers to the ability of the regulatory system to impose effective, proportionate, and dissuasive measures in cases of non-compliance. It includes the level of penalties, frequency of enforcement, transparency of decisions, and credibility of supervisory intervention. Sanctions are important not only as punitive instruments but also as mechanisms for shaping organizational behavior and increasing perceived regulatory risk.

6. Technical and Organizational Implementation (T)

Technical and organizational implementation concerns practical controls such as risk assessment, access control, incident response, backups, training, ISMS and privacy governance.

This variable represents the operational dimension of governance and determines whether formal obligations are translated into real resilience capacity.

7. Mathematical Formulation

The model conceptualizes digital trust as the result of the interaction between the four variables:

$$DT = N \times I \times S \times T$$

The multiplicative structure reflects systemic interdependence: weakness in one variable can substantially reduce overall digital trust, so legal precision alone cannot compensate for weak institutions or implementation.

8. Operational Indicators

Table 3. Operational Indicators Of The Digital Trust Governance Model

Variable	Operational Indicators
N - Normative Precision	clear incident reporting rules; defined management duties; DPO obligations; DPIA requirements; alignment with NIS2 and GDPR
I - Institutional Capacity	number of specialists; technical expertise; CSIRT coordination; supervisory independence; budgetary resources
S - Sanctioning Capability	level of penalties; frequency of enforcement; transparency of decisions; corrective measures; deterrent effect
T - Technical Implementation	ISMS adoption; incident response plans; access control; staff training; backup testing; privacy-by-design mechanisms

Source: Author's qualitative assessment framework based on cybersecurity governance maturity literature.

9. Illustrative Governance Assessment of Bosnia and Herzegovina

The framework is applied illustratively to BiH using qualitative evidence on regulatory alignment, institutional coordination, supervision and operational implementation.

Table 4. Illustrative Governance Maturity Assessment Of Bosnia And Herzegovina

Variable	Assessment	Score
Normative Precision (N)	Partial alignment with NIS2 and GDPR principles; moderate legal clarity	3
Institutional Capacity (I)	Fragmented institutional coordination and limited operational capacity	2
Sanctioning Capability (S)	Weak enforcement consistency and limited supervisory authority	2

Technical Implementation (T)	Uneven implementation maturity across sectors and institutions	2

Source: Author's qualitative assessment framework based on cybersecurity governance maturity literature.

Based on the proposed model: DT = N x I x S x T = 3 x 2 x 2 x 2 = 24.

The resulting DT score of 24 indicates moderate-to-low governance maturity, constrained by fragmented institutions, inconsistent enforcement and uneven implementation.

10. Application to Transitional Regulatory Systems

The model is especially relevant to transitional systems where legal reform may advance faster than institutional and technical capacity; in BiH, N is moderate while I, S and T remain comparatively weaker.

11. Comparative Governance Benchmarking

A qualitative benchmark compares BiH with Croatia and Germany using the same four dimensions to demonstrate the model's comparative applicability.

Table 5. Comparative Digital Trust Governance Assessment

Country	**N**	**I**	**S**	**T**	**Overall DT**
Bosnia and Herzegovina	3	2	2	2	24
Croatia	4	3	3	3	108
Germany	5	5	5	5	625

Source: Author's qualitative assessment framework based on cybersecurity governance maturity literature.

The illustration suggests that balanced legal clarity, institutional effectiveness, enforcement credibility and implementation capacity produce higher digital-trust maturity; BiH scores lower because of governance fragmentation and limited coordination.

12. Interpretation of the Multiplicative Governance Structure

The multiplicative model represents the systemic interdependence of governance dimensions. Unlike additive approaches, it assumes that a weak dimension can disproportionately reduce overall trust; strong legal harmonization cannot offset poor coordination, enforcement or implementation. This is consistent with resilience literature emphasizing interconnected organizational, institutional and technical capacities [22]. The formula is an analytical maturity framework, not a deterministic predictive instrument.

13. Theoretical and Practical Relevance

The model links regulatory effectiveness, cyber resilience, privacy governance and institutional accountability and can help policymakers and organizations identify priority weaknesses and move from formal compliance toward measurable governance effectiveness.

VII. DISCUSSION

1. Interpretation of Findings

The findings confirm an asymmetric BiH governance structure: privacy protection is relatively more consolidated, while cybersecurity remains institutionally fragmented and operationally uneven.

2. From Formal Compliance to Governance Maturity

Governance maturity differs from formal compliance because it requires practical capacity to prevent, detect, respond to and recover from incidents rather than merely maintaining policies and legal documents.

The $DT = N \times I \times S \times T$ model expresses this by making overall effectiveness sensitive to the weakest governance component.

3. Integration of Cybersecurity and Privacy Governance

Cybersecurity and privacy cannot be managed effectively as isolated domains. NIS2 strengthens resilience and management accountability, GDPR strengthens privacy accountability, and ISO/IEC 27001/27701 provide an operational bridge between them.

4. Role of Management Accountability

NIS2's emphasis on management bodies shifts cybersecurity toward strategic organizational accountability rather than a purely technical IT responsibility.

Management should approve policies, allocate resources, supervise implementation and ensure continuity and incident readiness. In BiH, clearer executive duties would strengthen both normative precision and implementation maturity.

5. Implications for Policy and Practice

The findings imply coordinated action by policymakers, regulators and organizations.

Priorities include clearer national responsibilities and NIS2 alignment, stronger supervisory and CSIRT cooperation, and organizational adoption of measurable ISO-based controls such as risk assessment, incident response, access management, training, backup testing and privacy by design.

6. Theoretical Contribution

The theoretical contribution is the treatment of digital trust as a measurable governance outcome shaped jointly by normative precision, institutional capacity, sanctioning capability and technical implementation.

7. Practical Contribution

Practically, the model supports regulatory planning, organizational audits, NIS2 readiness, ISO/IEC 27001 and ISO/IEC 27701 implementation assessment, and cyber-resilience benchmarking.

VIII. CONCLUSION

1. Main Findings

This paper examined integrated cybersecurity, privacy and digital trust governance through NIS2, ISO/IEC 27001 and ISO/IEC 27701 and showed that effectiveness depends on legal clarity, institutional capacity, enforcement credibility and technical implementation. For Bosnia and Herzegovina, the main asymmetry is stronger privacy governance alongside fragmented cyber governance, limited coordination and weak management accountability. The principal gaps concern CSIRT coordination, incident reporting, supervision and executive responsibility, indicating that implementation capacity and institutional resilience are as important as legal harmonization.

2. Contribution of the Study

The principal contribution is the Digital Trust Governance Model: $DT = N \times I \times S \times T$

3. Policy and Organizational Implications

BiH should move toward an integrated governance architecture aligned with European cybersecurity and privacy standards.

Organizations should complement legal alignment with measurable ISO/IEC 27001 and ISO/IEC 27701 controls for risk, incidents, access, continuity, training and privacy by design.

4. Limitations and Future Research

The study is limited by reliance on legal, institutional and public sources, which restricts direct measurement of operational capacity.

Because the model is conceptual and qualitative, future research should develop and test quantitative or expert-based indicators for each variable.

5. Final Remarks

Digital trust cannot be achieved through regulation alone; it requires coordinated institutions, credible enforcement, responsible management and implemented security and privacy controls.

The proposed model frames digital trust as the measurable result of integrated regulation, institutional capacity, enforcement credibility and technical implementation.

REFERENCES

[1]. European Parliament and Council, "Regulation (EU) 2016/679 of 27 April 2016 (General Data Protection Regulation)," Official Journal of the European Union, L 119, pp. 1–88, 2016.

[2]. European Parliament and Council, "Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)," Official Journal of the European Union, L 333, pp. 80–152, 2022.

[3]. International Organization for Standardization, ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems — Requirements. Geneva, Switzerland: ISO, 2022.

[4]. International Organization for Standardization, ISO/IEC 27701:2025, Information security, cybersecurity and privacy protection — Privacy information management systems — Requirements and guidance. Geneva, Switzerland: ISO, 2025.

[5]. National Institute of Standards and Technology, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29, 2024, doi: 10.6028/NIST.CSWP.29.

[6]. D. J. Solove, Understanding Privacy. Cambridge, MA, USA: Harvard University Press, 2008.

[7]. L. Floridi, "Soft ethics and the governance of the digital," Philosophy & Technology, vol. 31, no. 1, pp. 1–8, 2018, doi: 10.1007/s13347-018-0303-9.

[8]. R. Baldwin, M. Cave, and M. Lodge, Understanding Regulation: Theory, Strategy, and Practice, 2nd ed. Oxford, U.K.: Oxford University Press, 2012.

[9]. U. Beck, Risk Society: Towards a New Modernity. London, U.K.: Sage Publications, 1992.

[10]. Organisation for Economic Co-operation and Development, Risk and Regulatory Policy: Improving the Governance of Risk. Paris, France: OECD Publishing, 2010.

[11]. R. Pound, "Law in books and law in action," American Law Review, vol. 44, pp. 12–36, 1910.

[12]. European Parliament and Council, "Regulation (EU) 2022/2554 of 14 December 2022 on digital operational resilience for the financial sector (DORA)," Official Journal of the European Union, 2022.

[13]. European Union Agency for Cybersecurity, ENISA Threat Landscape 2023. Athens, Greece: ENISA, 2023.

[14]. European Union Agency for Cybersecurity, ENISA Threat Landscape 2024. Athens, Greece: ENISA, 2024.

[15]. European Union Agency for Cybersecurity, NIS360 2024: Cybersecurity Maturity and Criticality Assessment of NIS2 Sectors. Athens, Greece: ENISA, 2025.

[16]. European Union Agency for Cybersecurity, NIS2 Technical Implementation Guidance. Athens, Greece: ENISA, 2025.

[17]. European Commission, "Commission Implementing Regulation (EU) 2024/2690 of 17 October 2024 laying down rules for the application of Directive (EU) 2022/2555 as regards technical and methodological requirements of cybersecurity risk-management measures," Official Journal of the European Union, 2024.

[18]. European Cyber Security Organisation, ECSO White Paper: NIS2 Implementation. Brussels, Belgium: ECSO, 2025.

[19]. Organisation for Economic Co-operation and Development, Recommendation of the Council on Digital Security Risk Management. Paris, France: OECD, 2022.

[20]. Organisation for Economic Co-operation and Development, Digital Security and Resilience in Critical Infrastructure. Paris, France: OECD Publishing, 2023.
[21]. World Economic Forum, Global Cybersecurity Outlook 2024. Geneva, Switzerland: World Economic Forum, 2024.
[22]. R. von Solms and J. van Niekerk, “From information security to cyber security,” Computers & Security, vol. 38, pp. 97–102, 2013.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Reward System For Student Social Activity: Lntu Case Study

Iurii Lukianchuk
Department of Computer Science
Lutsk National Technical University
Lutsk, Ukraine
iuriilukianchuk87@gmail.com
ORCID: 0000-0001-9690-6197

Yurii Tulashvili
Department of Computer Science
Lutsk National Technical University
Lutsk, Ukraine
y.tulashvili@lutsk-ntu.com.ua
0000-0002-0780-9529

Lev Boyko
Department of Software Engineering
Lutsk National Technical University
Lutsk, Ukraine
lew.lazarus@gmail.com
0009-0001-0117-8551

Olga Nedilko
Department of Computer Science
Lutsk National Technical University
Lutsk, Ukraine
o.nedilko@lutsk-ntu.com.ua
0000-0002-7416-8899

Abstract
This paper addresses the problem of low social activity among students at higher education institutions and proposes a solution through a reward system implemented in a mobile application. Based on the analysis of global gamification practices in IT education, particularly at the University of Doha, the University of London, and the University of Northampton, it has been established that existing systems are limited to virtual rewards. The system proposed by the authors is the first among Ukrainian higher education institutions to combine the completion of social tasks on external networks, verification by an administrator, the accumulation of virtual points, and their exchange for the university's physical merchandise. The paper describes a pilot implementation among 100 students. Based on survey data from this pilot, theoretical projections are substantiated: an increase in student survey participation from 35% to 85-92% and a 300% increase in the university's social media presence. The prospects for scaling the system to 12,000 students are outlined.

Keywords: gamification, reward system, student social activity, mobile application, virtual points, higher education

I. INTRODUCTION

The modern development of information technologies creates new opportunities for increasing the efficiency of the educational process in higher education institutions. As researchers note [1, 2], special attention is drawn to gamification methods – the use of game mechanics in non-game contexts to increase user motivation and engagement. In higher education, gamification is used to enhance students' learning activities, increase class attendance, and improve academic performance [3]. According to the classical work of Karl Kapp [4], gamification of the educational process is based on elements such as points, badges, leaderboards, and levels, which together create a gaming experience for students. Leading universities around the world are implementing gamification elements into their educational programs. For example, the University of Doha (Qatar) is implementing a three-year research project to integrate gamification into programming education, using the Octalysis framework by Yu-kai Chou [5], which identifies eight key drivers of human motivation. The University of London uses the Level Up system, which allows students to earn experience points (XP) for completing educational activities, including watching video lectures, performing laboratory work, and passing tests, and displays their progress on the leaderboard. Research conducted at the University of Northampton confirms the positive impact of combining points, badges, and leaderboards on student attendance and academic performance. In addition, the work [6] presents a systematic literature review of gamification in educational

programs, confirming the growing interest in this topic. However, the analysis of global experience reveals a significant limitation: none of the existing systems provides for rewarding students with physical objects (merchandise). According to a systematic review, most studies focus on virtual rewards (points, badges, levels, leaderboards), which have a limited long-term motivational effect [7]. In addition, there are no examples of using gamification to increase students' social activity on external social networks, which is an important factor in the media presence of a modern university. It is also worth noting the work [8], in which the authors propose using artificial intelligence to gamify programming exercises, but this approach does not provide material rewards. This work aims to substantiate and implement a reward system in a mobile application to increase students' social activity at Lutsk National Technical University (LNTU) by combining virtual points with the option to redeem them for the university's physical merchandise.

II. RELATED WORK

Gamification in education has been widely studied. Kapp defined the core game elements applicable to non-game contexts. Chou introduced the Octalysis framework, which identifies eight human motivation drivers. More recent systematic reviews confirm that points, badges, and leaderboards remain the dominant gamification mechanisms. Comparative analysis of existing implementations at the University of Doha, the University of London, and the University of Northampton shows that none of these systems integrates material rewards (merchandise) or focuses specifically on students' social activity on external social networks. Montella et al. proposed an AI-powered gamification tool for programming exercises, but their approach lacks tangible incentives. The theoretical gap addressed by this paper is the lack of a closed motivation cycle (task → action → verification → accumulation → material reward) in existing higher education gamification systems. Self-determination theory suggests that external rewards can sometimes undermine intrinsic motivation; however, when rewards are perceived as informational rather than controlling, they can enhance engagement. The proposed system is designed with this distinction in mind.

III. METHODOLOGY

The analysis of LNTU students' social activity before the system's implementation revealed several problems. Firstly, student participation in university surveys is low, at about 35% of the total student population. Secondly, the university's presence on social networks, particularly Instagram and Facebook, is insufficient, as students rarely repost or share content about university life. Thirdly, there is no incentive system for students who actively promote the university. Fourthly, there is low student feedback through formal communication channels. Solving these problems is particularly relevant in the context of growing competition between higher education institutions for applicants. The proposed reward system is implemented as part of a mobile application developed using the cross-platform technology Flutter (Dart language). The application operates on iOS and Android platforms. The main components of the system are the task module, verification module, balance and store module, and notification module. The task module maintains a list of social tasks, each assigned a complexity coefficient: easy tasks receive a coefficient of 1, medium tasks receive a coefficient of 2, and difficult tasks receive a coefficient of 3. The base reward value is 10 conditional points, multiplied by the complexity coefficient to differentiate the reward based on the effort expended. Tasks are updated weekly (every 7 days), except for those with long execution periods. Task types include: liking, commenting, posting a post or story, reposting news, and participating in surveys. The verification module is designed to confirm task completion. After acting on a social network, the student uploads a screenshot or photo confirmation. Verification is performed manually by an administrator, after which the task status changes from "Under review" to "Completed," and points are credited to the user's balance. This approach to verification, although it requires human resources, ensures high reliability of results, unlike automatic systems, which can be deceived. The balance and store module displays the current balance of accumulated points. It includes a catalog of LNTU merchandise, which includes a mug, shopper, umbrella, notepad, and pen, each with a fixed cost in points. Exchange requests are accepted continuously, but physical merchandise issuance is scheduled once a month.

Such deferred material rewards create an additional motivational effect, as students can accumulate points throughout the month and receive desired items. The notification module provides push notifications that inform students when new tasks appear at the beginning of the week and when task status changes. The reward system's algorithm includes a sequence of steps that begins with user authentication in the application. After logging in, the student views the available tasks in the "In progress" category and selects one based on their preferences and capabilities. Next, the user performs the necessary action on an external social network, then uploads proof of completion as a screenshot or photo directly into the application. After that, the task receives the "Under review" status and waits for the administrator's verification. Upon successful verification, the corresponding number of points is credited to the user's balance, and the task status changes to "Completed." In case of rejection, the student can receive comments and try to complete the task again. Accumulated points can be redeemed in the application store for any product in the merchandise catalog. Ordered material rewards are received by students once a month at the university distribution point. Figure 1 shows the flowchart of the reward system.

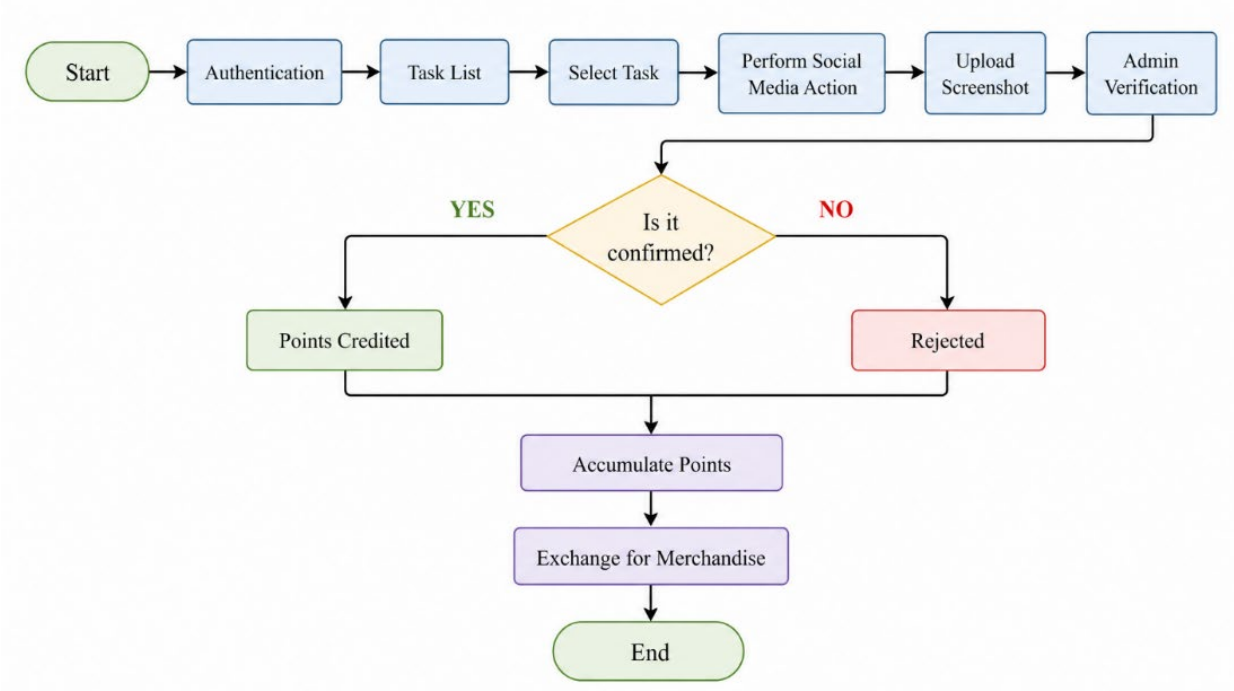


Figure 1. Reward system flowchart

To substantiate the novelty of the proposed system, a comparative analysis with existing solutions was conducted, and the results are summarized in Table I. As shown in the table, all analyzed systems use virtual points, badges, and levels, but none provide material rewards such as merchandise. The University of Doha and the University of London implement weekly task updates and leaderboards, but their systems do not include tasks on external social networks. The University of Northampton partially implements social tasks, but verification there is automatic, which does not guarantee reliable completion. The key difference of the LNTU system is the combination of virtual points with material rewards in the form of merchandise, which creates a closed motivation cycle: task → action → verification → accumulation → material reward. None of the analyzed higher education institutions applies a similar approach.

Table 1. Comparison of reward systems in higher education institutions

Characteristic	University of Doha	University of London	University of Northampton	LNTU (proposed)
Virtual points	Yes	Yes	Yes	Yes
Badges/levels	Yes	Yes	Yes	Yes
Leaderboard	Yes	Yes	Yes	In plans
Tasks on social networks	No	No	Partially	Yes
Completion verification	Automatic	Automatic	Automatic	Manual (administrator)
Material reward	No	No	No	Yes

A pilot implementation was conducted with 100 LNTU students in April 2026. Participation was voluntary and anonymous. Students were invited to use the mobile application and complete social tasks over four weeks. After this period, a survey was administered to collect feedback on perceived motivation, ease of use, and willingness to continue using the system. No control group was used in this pilot. Therefore, all reported results are theoretical projections based on survey responses and behavioral observations, not controlled experimental outcomes. The projections (85-92% survey participation, 300% increase in social media mentions) were derived from students' self-reported intentions and from extrapolation of task completion rates observed during the pilot. Ethical approval was obtained from the LNTU Institutional Review Board. All participants gave informed consent, and no personal identifying data were collected.

IV. RESULTS AND DISCUSSION

As of April 2026, the developed application is in pilot testing with 100 students participating. Based on survey responses from these 100 users, the following results are predicted after scaling the system to the entire university, where 12,000 students study. As shown in Figure 2, a significant increase in student participation in surveys is expected: from the current 35% to the predicted 85-92%, which is an increase of 50-57 percentage points. These projections are theoretical and require experimental validation with a control group.

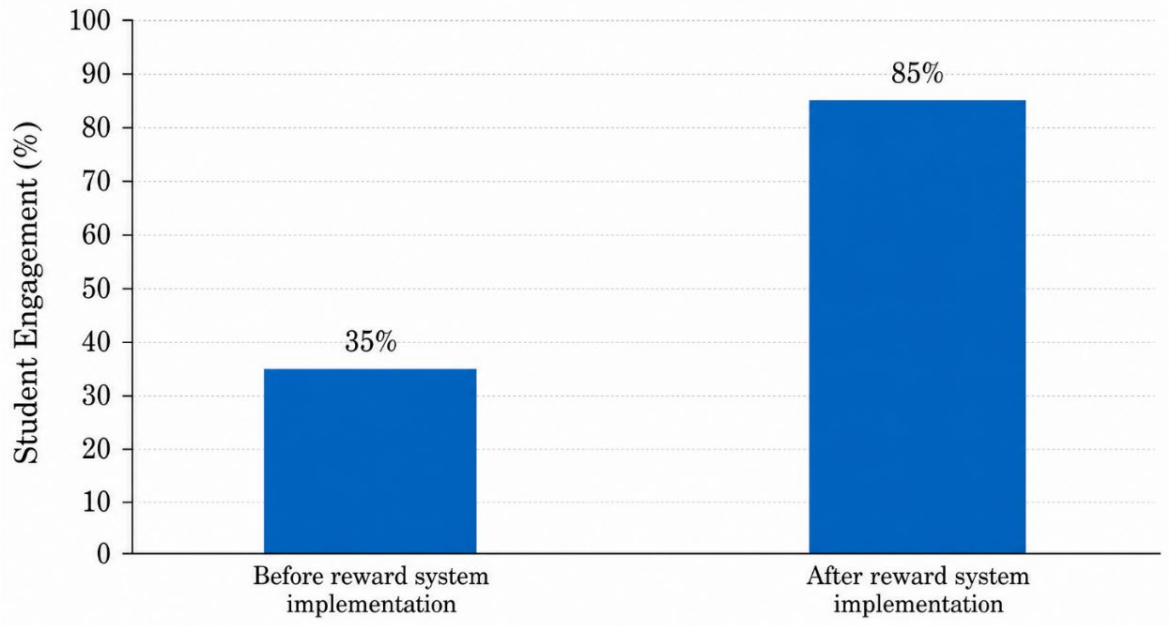


Figure 2. Expected growth in student survey participation.

In addition, an increase in the university's media presence on social networks is predicted: the number of LNTU mentions per month, according to Figure 3, should increase from about 50 to 200, i.e., by 300%. This projection is based on the observed frequency of task completion in the pilot group.

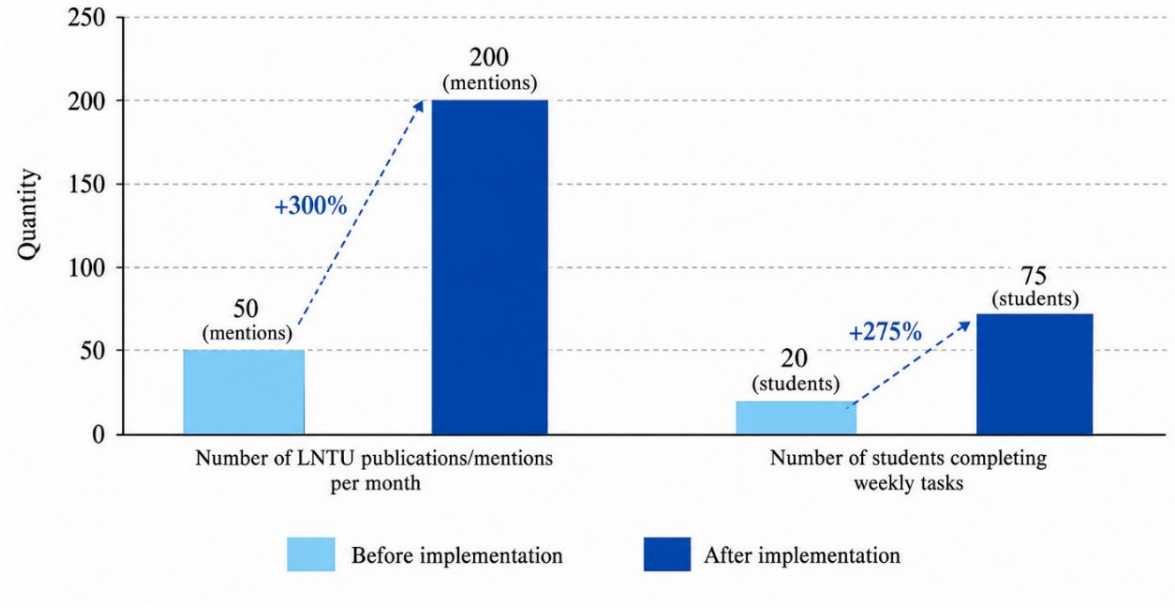


Figure 3. Expected growth in social media activity.

Also expected is an increase in the number of students who perform social tasks weekly, from 20 to 75 people. The economic efficiency of the proposed system is evident in the fact that the cost of merchandise (a mug costs about 150 UAH, a shopper – about 200 UAH) is lower than the cost of traditional advertising on social networks required to achieve the same level of audience coverage. This study has several limitations. First, the reported results are theoretical projections, not measured outcomes. Second, the pilot did not include a control group, so causal attribution of behavioral change to the reward system is not possible. Third, manual verification by an administrator may not scale to 12,000 students. Fourth, the long-term effect on intrinsic motivation was not assessed.

V. CONCLUSION

The article proposes and substantiates a reward system to increase students' social activity, implemented in a mobile application. The main results of the study are as follows. Firstly, an analysis of the global experience of gamification in IT education was conducted, using the University of Doha, the University of London, and the University of Northampton as examples, revealing the absence of systems that combine virtual points with material rewards such as merchandise. Secondly, for the first time at a higher education institution in Ukraine, a comprehensive approach to motivating students by completing tasks on external social networks, with administrator verification, accumulation of virtual points, and exchange for the university's physical merchandise is proposed. Thirdly, the architecture and algorithm of the reward system are presented, enabling the solution to scale from the current 100 students in the pilot group to 12,000 students. Fourthly, theoretical projections of key indicators are substantiated based on pilot survey data: participation in surveys may increase from 35% to 85-

92%, and the media presence of the university may increase by 300%. These projections require experimental validation. Further research will focus on automating task verification, using artificial intelligence for image recognition, integrating the system with the electronic hryvnia (e-hryvnia), and conducting a controlled experiment with a control group.

ACKNOWLEDGEMENT

The authors thank Lutsk National Technical University for supporting the pilot implementation of the reward system.

REFERENCES

[1]. Y. Tulashvili, I. Lukianchuk, V. Lishchyna, N. Lishchyna, "Program Implementation of Educational Electronic Resource for Inclusive Education of People with Visual Impairment," Lecture Notes on Data Engineering and Communications Technologies, vol. 180, pp. 251-260, 2023. DOI: 10.1007/978-3-031-36115-9_24.

[2]. R. Mellado, C. Cubillos, R. M. Vicari, G. Gasca-Hurtado, "Leveraging Gamification in ICT Education: Examining Gender Differences and Learning Outcomes in Programming Courses," Applied Sciences, vol. 14, no. 17, p. 7933, 2024.

[3]. V. Di Nardo, R. Fino, M. Fiore, G. Mignogna, M. Mongiello, G. Simeone, "Usage of Gamification Techniques in Software Engineering Education and Training: A Systematic Review," Computers, vol. 13, no. 8, p. 196, 2024.

[4]. K. M. Kapp, The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education. San Francisco: Pfeiffer, 2012.

[5]. Y. Chou, Actionable Gamification: Beyond Points, Badges, and Leaderboards. Octalysis Media, 2019.

[6]. M. Venter, L. De Wet, "Systematic Literature Review of Gamification Design in Higher Education Programming Courses: Methodological Rigor Exposed," Communications in Computer and Information Science, vol. 2250, pp. 249-268, 2025.

[7]. R. M. Ryan, E. L. Deci, Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York: Guilford Press, 2017.

[8]. R. Montella, C. G. De Vita, G. Mellone, T. Ciricillo et al., "GAMAI, an AI-Powered Programming Exercise Gamifier Tool," Communications in Computer and Information Science, vol. 2150, pp. 485-493, 2024.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Adaptive Lossless Compression for Event-Based Cameras

Prottasha D'Cruze
Dept of Mathematics and Computer Science
University of Lethbridge
Lethbridge, Canada
dcruze@uleth.ca
ORCID: 0009-0004-3318-3373

Howard Cheng
Dept of Mathematics and Computer Science
University of Lethbridge
Lethbridge, Canada
howard.cheng@uleth.ca
ORCID: 0000-0001-6689-4439

Abstract
We describe our recent work on improving lossless data compression performance for event-based cameras. Starting with the spike coding algorithm of Bi, Dong, Tian, and Huang, as well as the Motion Compensated Compression algorithm based on Optical Flow (MCCOF) of Bairagi and Cheng, this paper explores two directions to improve compression performance by adapting to the data being compressed. These include using K-means clustering to reduce spatial redundancy and applying adaptive event partitioning to the MCCOF algorithm. Preliminary experimental results on the DVS-PKU dataset show positive improvements for the MCCOF algorithm with adaptive partitioning in most cases, whereas K-means clustering underperforms. Ongoing research directions on cube-based motion compensation are also outlined.

Keywords: event-based cameras, data compression

I. INTRODUCTION

While conventional frame-based cameras capture video data by taking a series of pictures (or frames) in quick succession, event-based cameras operate in a very different manner. In event-based cameras, each pixel is an independent sensor that can detect significant changes in log-luminance and report them asynchronously [1]. Such changes are reported as an event containing a timestamp, the coordinates, and the polarity of the change (light-to-dark or dark-to-light). The exact log-luminance level is not reported. While this may provide fewer details about the scene, it allows the cameras to detect significant changes with very low latency and can generate events with sub-millisecond precision. Due to the low latency and high dynamic range of event-based cameras, real-time applications such as autonomous vehicle navigation can benefit from their use.

Event-based cameras naturally reduce the amount of data generated because pixels that do not change significantly do not generate any events. However, the high temporal resolution of these cameras means that a large volume of data can still be generated in an active scene. For example, a falling waterdrop can generate 46.2 million bytes of data over 6 seconds, representing almost 11.5 million events [2]. The transmission and, if needed, storage of these events present a difficult challenge. Compression is needed to handle this amount of data. In this work, we focus on lossless compression.

The events generated by these cameras are asynchronously ordered in time as sensors at various pixels report significant changes. As a result, the coordinates and polarity of consecutive events may not have a high spatial correlation. Conventional image and video compression algorithms that exploit spatial redundancies would need to be adapted to handle this type of data. Collecting the events into "frames" before processing would allow some of these redundancies to be exploited, at the expense of increasing latency of the compression algorithm.



Several algorithms have been developed for compressing data from event-based cameras. For example, Bi, Dong, Tian, and Huang proposed a spike-coding lossless compression algorithm [3], in which events are collected into "macro-cubes" temporally and "spike-cubes" spatially. Each spike-cube is encoded using one of two possible modes, depending on whether the events are spatially clustered or distributed in the cube. They later modified their algorithm to incorporate inter-cube prediction, leveraging temporal correlations among macro-cubes [4]. Such predictions are computationally costly and produced modest improvements in compression performance. Khan, Iqbal, and Martini proposed a time-aggregation-based algorithm that outperformed the spike-coding algorithm [5]. It works by aggregating events into event frames, which are encoded by conventional video compression algorithms. Unfortunately, this incurs a high computational cost and high latency. Other compression algorithms not specifically designed for event-based cameras can also be used [6], but they are not as effective.

Bairagi and Cheng proposed a new method to predict current events from previous events [7], [8]. They applied an event-based optical flow algorithm by Ridwan and Cheng [9], [10] to efficiently perform predictions in the Motion Compensated Compression algorithm based on Optical Flow (MCCOF). The PKU-DVS dataset [2] was used to evaluate the MCCOF algorithm, and it performed better than the spike-coding algorithm on most scenes.

In this paper, we report work on two different directions to improve the spike-coding and MCCOF algorithms. First, we examined the use of K-means clustering within each macro-cube in spike-coding for macro-cubes with more than one spatial cluster. Next, we introduce an adaptive version of the MCCOF algorithm, in which events in macro-cubes are partitioned based on the data. We will show that the adaptive version improves the compression performance of the MCCOF algorithm in many cases. Finally, we propose a possible cube-based motion compensation scheme, similar to that used in frame-based video compression algorithms, that shows promise for further exploration.

The paper is organized as follows. In Section II, background information and previous approaches for data compression for event-based cameras are given. In Section III, we describe our experiments with K-means clustering to improve spike coding. In Section IV, we introduce an adaptive version of the MCCOF algorithm. A cube-based motion compensation scheme is proposed in Section V. Concluding remarks and future directions are given in Section VI.

II. BACKGROUND

1. Event-Based Cameras

Event-based cameras are biologically inspired and have individual sensors for each pixel to detect significant log-luminance changes. When a sensor detects a significant change, it generates an event asynchronously, reporting the coordinates, the change $.(x, y, t, p)$, where $p \in \{-, +\}$. While frame-based conventional cameras capture frames regularly even when there are no changes, event-based cameras only generate events when there are significant changes. Only the polarity of the change is reported, but the magnitude is not. Event-based cameras have low latency and high dynamic range, making them suitable for applications that require real-time response under varying lighting conditions . The Address Event Data (AEDAT) format is used to store DVS event streams [11]. Event-based cameras, such as the DVS 128, DAVIS 240, and DAVIS 346 produced by iniVation, are used in this work [1].

Figure 1 shows the events captured by the Davis Dynamic Vision Sensor (DVS) [10]. The figure shows the events generated by a moving DVS camera on a static scene containing an apple. Events are generated at locations where significant changes in log-luminance occur. Typically, these events are associated with object boundaries, or, in this case, with the boundaries of the highlights on the apple's surface as well.

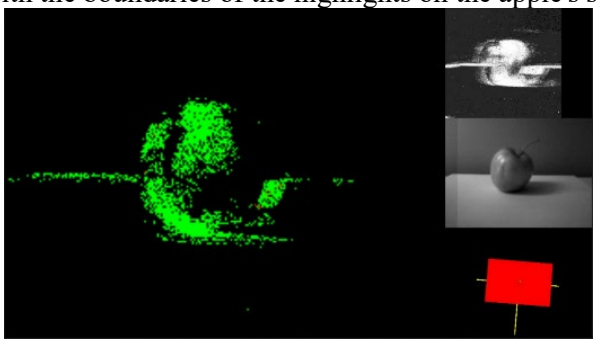


Figure 1. Events generated by a moving Dynamic Vision Sensor (DVS) on a static scene.

2. Previous Related Works

Bi, Dong, Tian, and Huang proposed a lossless predictive algorithm for DVS events (spikes) [3]. The sequence of events is partitioned into "macro-cubes" temporally adaptively, based on the length of time and number of events in the cubes. These macro-cubes are then further partitioned spatially into spike-cubes (Figure 2). By grouping events into temporally and spatially similar blocks, the algorithm can leverage redundancies among similar events to perform data compression. The locations and polarities of events are encoded by one of two methods, depending on whether the events in the spike-cubes are spatially clustered or distributed. The time-prior (TP) mode is used when events are spatially clustered and encodes the centroid location and each event's displacement from it. The address-prior (AP) mode is used when events are spatially distributed and encodes a binary map of event locations, followed by the remaining event details. The final encoding is done using a context-adaptive arithmetic coder [12].

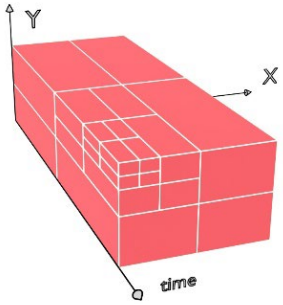


Figure 2. Partitioning events adaptively into spike-cubes.

Subsequently, the same authors [4] introduced reference AP and TP modes to exploit temporal correlations by predicting macro-cubes from previously encoded ones. The computational complexity of motion compensation is high, and, unfortunately, only a modest improvement in the compression ratio was achieved.

Bairagi and Cheng [7], [8] modified the spike coding algorithm [3] to incorporate motion compensation of events. They applied the optical flow algorithm of Ridwan and Cheng [9], [10] to efficiently generate motion predictions. Various strategies for encoding optical flow events (analogous to motion vectors in conventional video encoding) were studied, and the unpredicted residuals were then encoded using the original spike coding algorithm. The events were partitioned into spike-cubes in a non-adaptive manner to simplify the interaction between motion compensation and residual coding. Motion compensation was performed adaptively for each spike-cube only when it improved compression performance. The resulting algorithm, called the Motion-Compensated Compression Algorithm based on Optical Flow (MCCOF), outperformed the original spike-coding algorithm across most scenes in the PKU-DVS dataset. The architecture of the MCCOF algorithm is shown in Figure 3.

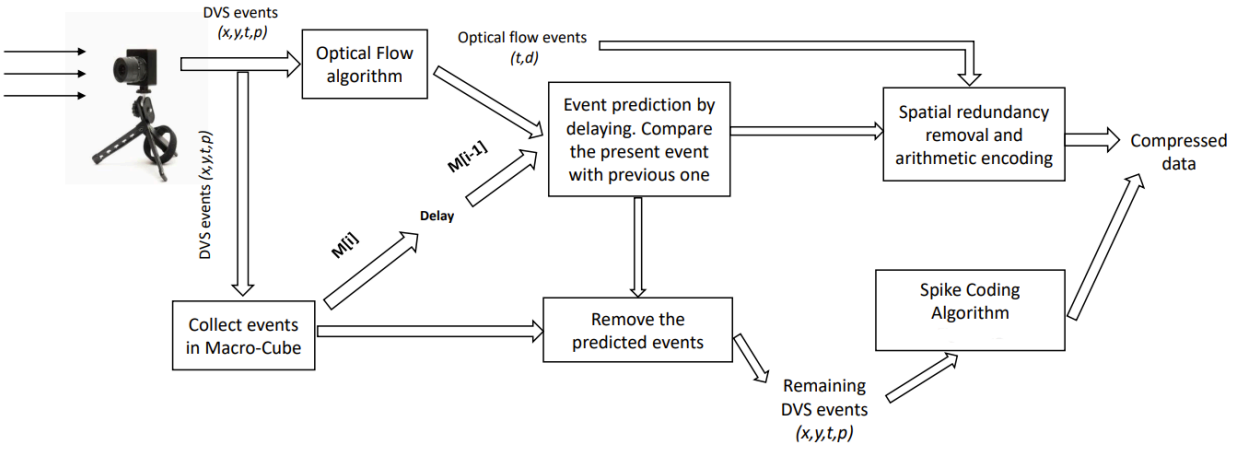


Figure 3. Overall architecture of the Motion Compensated Compression Algorithm based on Optical Flow (MCCOF).

III. K-MEANS CLUSTERING

In the spike coding algorithm of Bi, Dong, Tian, and Huang [3], two modes are used to encode spike-cubes depending on whether the events in that spike-cube are spatially clustered or distributed. In the former case, the time-prior (TP) mode is used. In the TP mode, the spatial centroid of all the events in that spike cube is computed, and the location is encoded. The events in that spike-cube are then temporally ordered, and the spatial displacements, the time difference from the previous event, and the polarity of each event are encoded. When most of the events are clustered around the centroid, this can result in significant compression. On the other hand, the address-prior (AP) mode is suitable when the events are distributed spatially throughout the spike-cube.

Events in each spike-cube are not always either in one single cluster or completely distributed. The TP mode could potentially be generalized to support multiple clusters. At the same time, however, the adaptive partitioning of events into spike-cubes may already form spike-cubes that do not have more than one cluster. We now examine this aspect in more detail. Instead of partitioning the temporal macro-cubes into spike-cubes, we propose applying K-means clustering [13] to each macro-cube to group events into a small number of spatial clusters. The elbow method is used to determine the number of clusters present in each spike-cube [14], and a sequential K-means algorithm is used to perform incremental clustering as events arrive. Once the clusters are identified, each cluster is encoded in the same way as in the original TP mode—the spatial coordinates of the

centroid are encoded, followed by the spatial displacement of each event from the centroid. The timestamps and polarities of the events are encoded as before.

Experimental results on the PKU-DVS dataset [2] are given in Table 1, comparing the compression ratios achieved by the proposed K-means compression algorithm and the spike coding algorithm. Experimental results show that the K-means compression algorithm generally underperforms, sometimes by a significant margin. However, there is one scene (night traffic) in which the K-means algorithm performs better. In that scene, several fixed objects (e.g., streetlights) consistently generate events at the same locations. On the other hand, in scenes where the K-means compression algorithm performs particularly poorly (e.g., fluorescent and intersection), the events are sparse and not spatially clustered in the scene at all. Although the fluorescent scene also has a fixed fluorescent light, many events are randomly distributed throughout the scene due to noise. In particular, random noise that can appear throughout the scene is forced to belong to certain clusters in this approach, and that leads to overall decreased performance. These results confirm that the adaptive partitioning of events into spike-cubes in the spike coding algorithm generally localizes events to single clusters. Still, in some rare cases, it can be beneficial to consider alternatives.

Table 1. Compression ratios from the spike coding algorithm and the K-means algorithm.

Scene	**Spike Coding**	**K-means**
waterdrop	**59.59**	50.51
fluorescent	**33.67**	10.14
lighter	**20.38**	9.08
football	**18.82**	11.79
jump	**8.97**	5.60
game	**4.50**	4.34
pendulum	**4.44**	3.00
intersection	**62.20**	11.86
pedestrians	**14.25**	5.44
daytime-traffic1	**11.13**	6.54
daytime-traffic2	**6.49**	3.97
night-roadside	**5.44**	4.55
night-traffic	3.87	**4.18**

IV. ADAPTIVE PARTITIONING IN THE MCCOF ALGORITHM

The Motion Compensated Compression Algorithm based on Optical Flow (MCCOF) [7], [8] uses optical flow calculations to perform motion compensation, thereby improving compression performance. Unlike conventional frame-based video encoding algorithms, motion compensation is performed for each pair of events rather than for each pair of coefficient blocks [6]. Optical flow events serve as motion vectors and are encoded, and any unpredicted events (the residual) are encoded using the spike coding algorithm. Due to how the optical flow prediction and spike coding algorithms were integrated into the MCCOF algorithm, the spike coding algorithm used only fixed-size macro-cubes and spike-cubes. In the original MCCOF algorithm, each macro-cube contains a fixed number of events, and each spike-cube has the same spatial dimensions arranged in a regular grid. As a result, some sparse macro-cubes may represent a long elapsed time, and some spike-cubes

may have very few events, leading to ineffective compression. Instead, we adapted our algorithm to split macro-cubes and spike-cubes adaptively, without using a tree structure, as in the spike-coding algorithm. Otherwise, the tree structure would also have to be encoded and transmitted for every macro-cube.

In our proposed adaptive MCCOF algorithm, the events are first spatially partitioned by a regular fixed-sizegrid into 10 0 $blocks$ ($S = 10$ [8]). For each spatial partition, we collect events temporally adaptively into spike-cubes, by a threshold on either the elapsed time between the first and the last event in that spike-cube, or the number of events in that spike-cube (Figure 4). The algorithm determines the bit rate achieved with different partitions of spike cubes and adaptively chooses the one that yields the lowest bit rate. Notice that the decoder only needs to be informed when each cube ends in the event stream, which is required in the original MCCOF algorithm. No additional data needs to be encoded to describe the partitioning. Using a fixed spatial grid requires no additional bit budget to encode the spatial partitioning of events. In the original MCCOF algorithm, predicted events are removed, and the spike-coding algorithm encodes only the residuals. As a result, the residual events coded by the spike coding algorithm are much sparser. The proposed adaptive partitioning strikes a trade-off between ensuring a sufficient number of events in a spike-cube to exploit redundancies and achieving a spike-cube with a long elapsed time, which may reduce temporal redundancies. Table 2 presents experimental results for the Adaptive MCCOF algorithm, compared with spike coding and the original MCCOF algorithm. We can see that in most cases, the adaptive MCCOF performs better than the original MCCOF and spike coding algorithm. In cases in which the spike coding algorithm performs better than the original MCCOF algorithm (waterdrop, fluorescent, and football), the Adaptive MCCOF algorithm improved MCCOF to surpass the original spike coding algorithm in all scenes except for the waterdrop scene. There are a few cases where the Adaptive MCCOF algorithm performs slightly worse than the original, and in the pedestrian scene, it performs significantly worse. The reason is that the original MCCOF algorithm selects the optimal spike-cube size spatially and temporally for the data set, even though the size is fixed across the entire event stream. On the other hand, the Adaptive MCCOF selects a single fixed spatial size and adapts only temporally within the event stream. This choice was deliberate to avoid a tree structure for representing the partitioning, but it sometimes performs slightly worse, as seen in our experimental results. If desired, one can use the Adaptive MCCOF algorithm to encode the data multiple times with different fixed spatial sizes and choose the one that yields the best performance; with that change, the Adaptive MCCOF can never underperform compared to the original MCCOF algorithm. However, the computational cost of compression would increase.

In the case of the pedestrian scene, the chosen spatial size was misaligned with the objects' sizes. By choosing different fixed sizes for the spatial partitioning of spike-cubes, the performance of the Adaptive MCCOF approaches that of the MCCOF Algorithm. Further investigation is needed to address this drawback and enable some adaptation of spatial partitioning to spike-cubes without introducing significant overhead for encoding the tree structure. It may be possible to use the spatial structure of one macro-cube to predict the spatial structure of the next one.

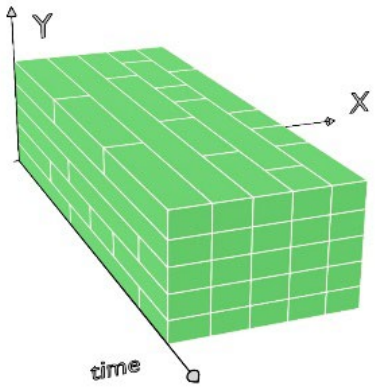


Figure 4. Partitioning events into spike-cubes in adaptive MCCOF.

Table 2. Compression Ratio from the Adaptive MCCOF algorithm compared to spike coding and the MCCOF algorithm.

Scene	Spike Coding	MCCOF	Adaptive MCCOF
waterdrop	**59.59**	54.74	57.34
fluorescent	33.67	32.14	**36.01**
lighter	20.38	**22.16**	21.51
football	18.82	18.44	**20.10**
jump	8.97	9.86	**9.87**
game	4.50	5.25	**5.26**
pendulum	4.44	**4.93**	4.89
intersection	62.20	76.93	**77.95**
pedestrians	14.25	**15.98**	8.96
daytime-traffic1	11.13	**12.99**	12.98
daytime-traffic2	6.49	**6.57**	6.51
night-roadside	5.44	6.34	**6.48**
night-traffic	3.87	4.71	**4.91**

V. CUBE-BASED MOTION COMPENSATION

Dong, Bi, Tian, and Huang [4] proposed the reference TP and reference AP modes to predict cube events, thereby further improving compression performance. For each spike-cube, a search is conducted to find the best motion vector that best matches the current spike-cube of events with a previous spike-cube, and the prediction residuals are encoded. The computational cost is high (three times that of the original spike coding algorithm [3]), and the improvement in compression performance is modest.

The MCCOF algorithm uses a highly efficient optical flow method to compute motion vectors quickly. These motion vectors can only be in one of eight possible compass directions, each of length 1. Still, the restriction allows the motion vectors to be computed quickly, and the compression performance of the MCCOF algorithm over spike coding justified this restriction. In addition, the optical flow algorithm [9], [10] is based on tracking single events, predicting only one event from another reference event. In the MCCOF and Adaptive MCCOF algorithms, each motion vector is used only to predict one other event.

While there are already significant improvements in compression performance over the spike coding algorithm, a cube-based motion compensation scheme can use a single motion vector for each spike-cube to predict a spike-cube of events, reducing the number of motion vectors to encode. We propose two approaches to incorporating cube-based motion compensation into existing compression algorithms.

In the first direction, we can apply the optical flow algorithm of Ridwan and Cheng [9], [10] to efficiently compute the motion vectors in the spike coding algorithm [4]. The motion vectors in the same spike-cube are collected, and the majority direction is used as the motion vector for the entire spike-cube. Prediction residuals are encoded in reference TP and AP mode as before. Preliminary results show promise for significantly reducing computational time while maintaining the improvements observed previously in [4]. We are investigating the possibility of allowing our optical flow algorithm to search over a wider range of motion vectors, trading off more accurate predictions for higher computational costs.

In a second direction, we would like to explore incorporating the same cube-based motion-vector computation into the MCCOF and Adaptive MCCOF algorithms. The results of this exploration will be forthcoming.

VI. CONCLUSIONS

In this paper, we examined two possible directions for improving the compression performance of the spike coding and MCCOF algorithms. The PKU-DVS dataset was used to evaluate the effectiveness of these improvements. First, we explored using K-means clustering to adapt the spike coding algorithm to handle better spike-cubes that may contain multiple spatial clusters. Experimental results show that, in general, the original spike-cube partitioning algorithm already yields localized clusters, and K-means clustering underperforms in all but one case. Further investigations may be warranted to understand that case. Next, we introduced adaptive partitioning to the MCCOF algorithm. The events are partitioned non-adaptively spatially but adaptively temporally to exploit spatial redundancies in the spike-cube and achieve optimal compression. The adaptive MCCOF algorithm outperforms the original MCCOF algorithm and the spike-coding algorithm in most cases.

Finally, we proposed two methods to incorporate cube-based motion compensation to further improve the performance of spike-coding and MCCOF, both in terms of compression performance and computational cost. These directions are currently being explored, and we expect the results of these investigations to appear in a forthcoming paper.

ACKNOWLEDGEMENT

This work was supported by a Chinook Summer Research Award at the University of Lethbridge.

REFERENCES

[1]. G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, "Event-Based Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, 2022, doi: 10.1109/TPAMI.2020.3008413.

[2]. National Engineering Laboratory for Video Technology, Peking University, "PKU-DVS Dataset," Jul. 2022.

[3]. Z. Bi, S. Dong, Y. Tian, and T. Huang, "Spike Coding for Dynamic Vision Sensors," in *2018 Data Compression Conference*, 2018, pp. 117–126. doi: 10.1109/DCC.2018.00020.

[4]. S. Dong, Z. Bi, Y. Tian, and T. Huang, "Spike Coding for Dynamic Vision Sensor in Intelligent Driving," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 60–71, 2019, doi: 10.1109/JIOT.2018.2872984.

[5]. N. Khan, K. Iqbal, and M. G. Martini, "Time-Aggregation-Based Lossless Video Encoding for Neuromorphic Vision Sensor Data," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 596–609, 2021, doi: 10.1109/JIOT.2020.3007866.

[6]. D. Salomon, *Data Compression: The Complete Reference*, 4th ed. Springer, 2007.

[7]. A. K. Bairagi, "Motion Compensated Compression for Event-based Cameras," M.Sc. thesis, University of Lethbridge, 2022.

[8]. A. K. Bairagi and H. Cheng, "Motion Compensated Compression for Event-Based Cameras," in *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, IEEE, Aug. 2025, pp. 1–6. doi: 10.1109/ACDSA65407.2025.11165984.

[9]. I. Ridwan, "Looming object detection with event-based cameras," M.Sc. thesis, University of Lethbridge, 2017.

[10]. I. Ridwan and H. Cheng, "An Event-Based Optical Flow Algorithm for Dynamic Vision Sensors," in *Image Analysis and Recognition (ICIAR) 2017*, 2017, pp. 182–189.

[11]. iniVation, "AEDAT File Format," Feb. 2022.

[12]. D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 620–636, 2003, doi: 10.1109/TCSVT.2003.815173.

[13]. J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Appl. Stat.*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.

[14]. B. Kamranian and H. Cheng, "Event-Based Looming Objects Detection," 2022, pp. 82–95. doi: 10.1007/978-3-030-96878-6_8.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Artificial Intelligence in Online Gambling and Betting Systems: A Knowledge Representation Based Use Case

Trajance Angelovski
Faculty of Computer Science
University of Skopje
Skopje, Republic of North Macedonia
trajance_angelovski@hotmail.com

Toni Malinovski
Faculty of Computer Science
University of Skopje
Skopje, Republic of North Macedonia
t.malinovski@utms.edu.mk

Sime Arsenovski
Faculty of Computer Science
University of Skopje
Skopje, Republic of North Macedonia
s.arsenovski@utms.edu.mk

Miodrag Cekikj
Faculty of Computer Science
University of Skopje
Skopje, Republic of North Macedonia
m.cekikj@utms.edu.mk

Abstract
This study investigates the application of knowledge-based Artificial Intelligence (AI) in online gambling and betting systems through a structured use-case approach. Focusing on semantic networks, frames, scripts, and production rules, the study demonstrates how these knowledge representation techniques can enhance system reliability, risk management, and decision transparency in regulated gambling environments. Each technique is analyzed according to its functional role, strengths, and limitations, and its combined use is examined through a comparative and integrative perspective. The findings indicate that integrating multiple knowledge representation techniques provides complementary capabilities that any single technique cannot achieve. The proposed approach highlights the continuing relevance of symbolic and hybrid AI solutions for explainable, ethical, and compliant online gambling platforms.

Keywords: AI, knowledge representation, online gambling, betting systems

I. INTRODUCTION

AI plays an increasingly important role in modern digital systems by enabling automated reasoning, decision-making, and pattern recognition [1], [2]. A fundamental component of intelligent behavior in AI is knowledge representation [3], which provides formal mechanisms for structuring and reasoning over domain knowledge in a way that is understandable and usable by machines. Recent research emphasizes that effective knowledge representation remains essential for building AI systems that move beyond raw data processing and support explainable, consistent, and auditable decision-making, particularly in complex and high-risk domains [4], [5]. The online gambling and betting industry represents a domain where AI technologies are widely adopted due to the availability of large-scale data, real-time operational requirements, and complex risk management challenges [6], [7]. AI-driven solutions can be applied for odds calculation, user profiling, fraud detection, regulatory compliance, and responsible gambling interventions [7]. However, studies indicate that purely data-driven approaches, such as machine learning models, do not always produce sufficient or reliable results in highly regulated environments that require transparency, interpretability, and ethical accountability [8]. Limitations related to explainability, data bias, and incomplete or dynamic user data highlight the need for complementary AI approaches and integration with established practices and human decisions [9]. Knowledge-based AI approaches, implemented through knowledge representation techniques, offer a structured and interpretable foundation for addressing these challenges. Techniques such as semantic networks, frames, scripts, and production rules enable explicit modeling of domain concepts, relationships, processes, and expert

decision logic [10]. Recent literature shows that such symbolic and rule-based approaches are particularly suitable for online gambling systems, where business rules, regulatory constraints, and behavioral patterns must be represented in a transparent, controllable, and verifiable manner, especially when supporting compliance and responsible gambling mechanisms [7], [8], [11], [12]. This study addresses the limitations of purely data-driven AI in online gambling systems, particularly regarding transparency and regulatory alignment. It examines how knowledge representation techniques, such as semantic networks, frames, scripts, and production rules, provide structured and interpretable decision support in a realistic use case. The study highlights their role as complementary mechanisms to existing AI solutions, improving system reliability, risk management, and compliance.

II. RELATED WORK

Recent research highlights that knowledge representation remains a foundational component of Artificial Intelligence, particularly in domains where explainability, accountability, and structured reasoning are required [10], [13]. While modern AI research is often dominated by machine learning and statistical models, several studies emphasize that symbolic, knowledge-based AI approaches continue to play a critical role in enabling transparent decision-making, semantic reasoning, and system interpretability [10], [14]. In regulated and high-risk environments, knowledge representation techniques provide mechanisms that allow AI systems to model explicit domain knowledge and support auditable reasoning processes [5], [13]. The application of AI in the online gambling and betting sector has expanded significantly, driven by large-scale user data, real-time analytics, and competitive market pressures. Recent empirical studies demonstrate that AI-based systems are widely used for player behavior analysis, fraud detection, personalization, and early identification of gambling-related risks [15], [16], [17]. Machine learning techniques, in particular, have shown strong performance in detecting anomalous behavior patterns and predicting at-risk players. However, contemporary research also notes that these models often lack sufficient transparency and may be difficult to align with regulatory requirements without additional control mechanisms [5], [18], [19]. To address these limitations, recent literature increasingly supports hybrid AI approaches that combine data-driven models with symbolic and rule-based knowledge representation. Such approaches allow system designers to encode business logic, regulatory constraints, and ethical safeguards directly into AI systems [19]. Among the established knowledge representation techniques, semantic networks are commonly used to model relationships among entities, which can be players, betting events, risks, and outcomes, enabling structured reasoning over interconnected concepts [20], [21], [22]. Frames extend this capability by representing entities with well-defined attributes, default values, and procedural behaviors, supporting consistency and reuse across system components [23]. Scripts have been applied to model typical sequences of actions, such as the lifecycle of an online betting session. At the same time, production rules remain a practical technique for encoding expert knowledge and decision logic in an interpretable deterministic form [24]. Recent studies emphasize that these techniques remain highly relevant when designing AI systems that must balance operational efficiency with transparency and ethical compliance [24], [25]. Although these approaches are well established, their systematic use in modern online gambling platforms remains an active area of research. Most studies focus on isolated AI components, with fewer exploring how knowledge representation can be integrated into coherent system architectures alongside analytical and predictive models. This study addresses that gap by examining semantic networks, frames, scripts, and production rules as practical tools within AI-enabled gambling systems. A representative use case shows how structured knowledge representation can be incorporated into real-world system design rather than treated purely as theoretical.

III. AI KNOWLEDGE REPRESENTATION–BASED USE CASE

This study follows a conceptual, use-case-oriented methodology to examine the application of knowledge representation techniques in AI-enabled online gambling and betting systems [1], [2], [4], [10], [13]. A real-world online gambling platform operating in the Republic of North Macedonia, covering both online and betting

shop activities, serves as the use case, reflecting key elements such as players, events, odds, accounts, risk indicators, and regulatory constraints. The methodology demonstrates how structured knowledge representation is applied to systematically model system behavior, risk-related processes, and decision-making structures. The analysis was conducted from August 2025 to April 2026 to capture typical system activities and interaction scenarios. The study examines four established knowledge representation techniques: semantic networks, frames, scripts, and production rules [1], [3], [4], [13]. These AI-enhanced techniques were selected for their interpretability and practical applicability in regulated environments. Each technique is analyzed according to its functional role within the use case: semantic networks for modeling relationships among core concepts, frames for structuring system entities and their attributes, scripts for representing typical sequences of betting-related activities, and production rules for encoding decision logic and policy constraints. Together, these methods provide complementary perspectives on how symbolic AI can enhance system transparency and control, thereby complementing human discussion in daily operations. Rather than evaluating predictive accuracy or algorithmic performance, the methodology emphasizes qualitative modeling and structural analysis. By examining how each representation technique contributes to specific system functions within the same use case context, the study highlights their combined value in supporting explainable, robust, and ethically aligned AI behavior, particularly when used alongside data-driven approaches.

1. Semantic Networks

Semantic networks are used as a core knowledge representation technique for modeling the conceptual structure of an AI-enabled online gambling and betting system. Their graph-based structure enables transparent and interpretable representation of entities and relationships, organizing domain knowledge such as players, betting activities, risk factors, and responsible gambling mechanisms [20], [21]. They define explicit relationships (edges) between entities (nodes), capturing system interactions and supporting reasoning about betting actions, risk levels, and policies, thereby forming a foundation for higher-level decision support. Figure 1 illustrates the constructed semantic network for the use case, showing the modeled entities, their labeled relationships, and key interaction dependencies.

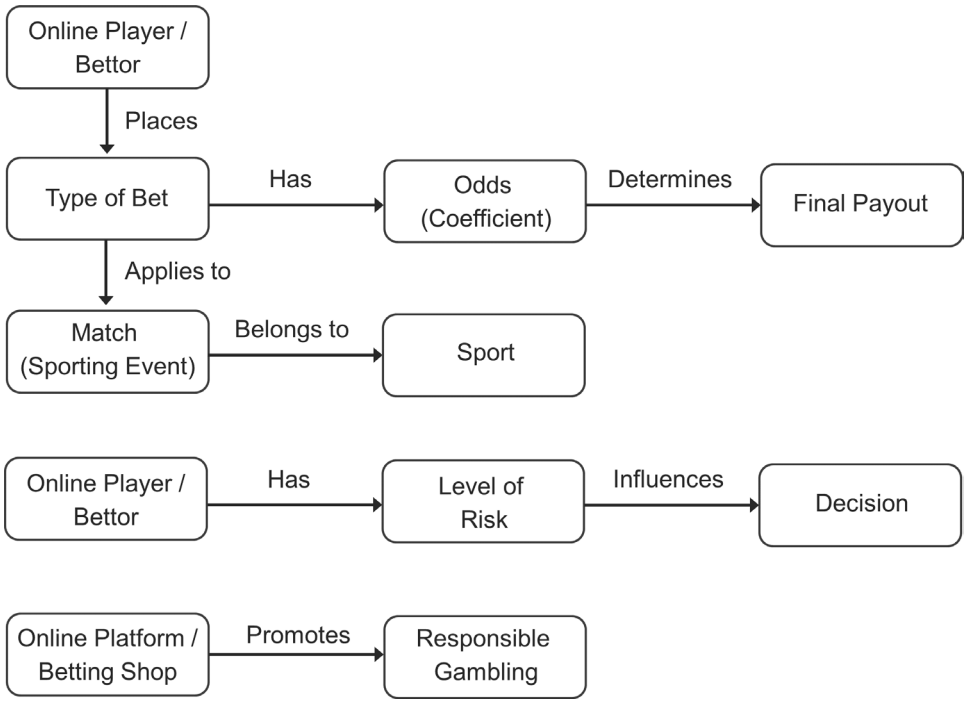


Figure 1. Semantic network representing core concepts and relationships in an online gambling and betting system

2. Frame-Based Representation

In this study, frames are used as a practical knowledge representation mechanism for modeling structured entities in an AI-enabled online gambling and betting system. Rather than focusing on theoretical classification, they are applied to organize entity-specific information in a clear, consistent, and controllable manner within the use case context. Frames represent core entities such as players, betting tickets, matches, and platform components by grouping related attributes and constraints into structured units. Each frame consists of slots for relevant attributes, including values, defaults, and optional procedural logic. This supports systematic knowledge storage and facilitates validation, updates, and analysis of individual entities [1], [3], [4], [13], [26], while containing information required to support risk assessment, betting eligibility checks, and compliance with responsible gambling policies.

To illustrate the use of frames in the proposed system, a Player / Bettor frame is defined using attributes directly derived from the operational requirements of an online gambling platform (Table 1).

Table 1. Player / Bettor Frame

Attribute	**Value**
Type	User (online, betting shop, or combined)
ID	Unique player identifier
Full Name	Text
Age	Numeric value
Account Balance	Real number (currency-dependent)
Risk Level	{low, medium, high} (default: medium)
Betting History	List of past bets with timestamps and outcomes
Betting Limit	Real number
Account Status	{active, suspended} (default: active)

3. Script-Based Representation

Scripts are used to model recurring interaction processes within an AI-enabled online gambling and betting system [23]. Unlike semantic networks and frames, which focus on entities and their attributes, scripts represent the ordered sequence of actions that typically occur during user interaction with the platform. This makes scripts particularly suitable for capturing behavioral flows and system dynamics in the use case. Within the proposed system, scripts are employed to describe the standard lifecycle of an online betting session. The script is triggered when a user intends to place a bet and proceeds through a predefined sequence of activities that reflects how users generally interact with the platform. Required input data includes bet amount, selected bet type, odds, time of placement, outcome, and user betting history. Before any evaluation or decision is produced, the script includes validation steps to ensure that the provided data are complete, realistic, and free of extreme or inconsistent values. The betting script in the use case consists of the following conceptual stages:

1. **Entry conditions**: Situations that must be fulfilled before the events in the script can take place. For example, for the "Online Betting" script: "The user wants to place a bet" and "The match meets the conditions that allow the user to place a bet."
2. **Props:** Objects used in the sequence of events. For the "Online Betting" script: odds, stake, type of game, time before the start, end time, result, comparison of data to determine whether there is a win or not, etc.
3. **Roles:** Performed by the participants. For the "Online Betting" script: user, match, result, etc.
4. **Scenes:** A real sequence of events. For the "Online Betting/Gambling" script, scenes may include:
   - Logging into the online betting platform
   - Choosing a game
   - Placing a stake
   - Type of game/way of playing
   - Win / loss
   - Payout
5. **Results:** Conditions that are fulfilled after the game is completed.

This structured sequence allows the system to reason about ongoing or incomplete interactions, such as detecting unfinished bets or abnormal betting behavior, and to maintain consistency in process execution. The AI integration enables fast and more accurate analysis of betting statistics and risk, far beyond what humans can process at the same speed. It helps detect fake accounts used for fraud and money laundering, provides 24/7 customer support through chatbots, and supports responsible gambling by monitoring player behavior, issuing warnings or restrictions when risky patterns appear, and reducing promotions for vulnerable players.

4. Production Rules

Production rules are used to represent decision logic and policy constraints in an AI-enabled online gambling and betting system. They provide an interpretable mechanism for enforcing decisions related to risk management, bonuses, account restrictions, and responsible gambling. Expressed in IF–THEN form, they link system actions to clearly defined conditions [10], [24].

In the proposed use case, they operate on data from other representations such as player frames, risk indicators, and betting history. When conditions are satisfied, corresponding actions are executed automatically, ensuring consistent and transparent behavior. This makes them well-suited to regulated environments where decision logic must be traceable and auditable. The following is an example of a risk management rule applied in the system:

IF (player.risk_level == high) AND (loss_duration == prolonged) THEN trigger warning AND prepare account restriction

Production rules are also used to model bonus eligibility logic, such as:

- Single-event ticket:

```
IF (bonus_active == true) AND
  (bet_odds >= 1.5) AND
  (total_turnover >= bonus_amount)
THEN bonus_status = completed
    allow_withdrawal = true
```

ELSE bonus_status = not_completed
allow_withdrawal = false

- Multi-event ticket:

IF (bonus_active == true) AND
(number_of_events >= 3) AND
(minimum_event_odds >= 1.45)
THEN ticket_qualifies_for_bonus = true
ELSE ticket_qualifies_for_bonus = false

These examples illustrate how expert knowledge and business rules can be encoded directly into the system, ensuring consistent enforcement of platform policies.

IV. COMPARISON AND INTEGRATION OF AI KNOWLEDGE REPRESENTATION TECHNIQUES

Based on the implementation findings, the AI knowledge representation techniques show complementary strengths individually and collectively, with key system functionality emerging most effectively when they are used in a coordinated rather than isolated manner.

5. Comparison of AI Knowledge Representation Techniques

Table 2 summarizes the key characteristics of the four knowledge representation techniques as applied in the proposed use case. The qualitative comparison is derived from their functional roles within the online gambling and betting system and reflects their respective strengths and limitations.

Table 2. Comparison of AI Knowledge Representation Techniques

Feature	Semantic Networks	Frames	Scripts	Production Rules
Type of Knowledge	Declarative	Declarative and Procedural	Declarative (process-oriented)	Procedural
Primary Structure	Graph of concepts and relations	Structured entities with attributes	Ordered sequence of actions	IF–THEN rules
Main Purpose	Represent relationships between concepts	Model system entities and states	Model user and system processes	Encode decision logic and policies
Flexibility	High	Moderate	Moderate	Variable
Reasoning Capability	Inheritance-based inference	Slot inheritance and defaults	Sequential reasoning	Logical rule evaluation
Ease of Modification	More complex due to interconnections	Relatively easy via slot updates	Requires script restructuring	Easy through rule addition/removal

Feature	Semantic Networks	Frames	Scripts	Production Rules
Suitability for Gambling Use Case	Modeling entities, risk links, and relationships	Managing player profiles and system data	Capturing betting session flow	Enforcing bonuses, limits, and compliance

This comparison shows that each technique addresses a different aspect of system intelligence. Semantic networks emphasize relationships; frame-structure entities and data; scripts organize behavioral flows; and production rules control decision logic.

6. Integration of AI Knowledge Representation Techniques in the Use Case

The analysis demonstrates that no single knowledge representation technique is sufficient to model the full complexity of an AI-enabled online gambling and betting system. Instead, the strongest results emerge when the techniques are combined into an integrated framework, where each technique complements the others [1], [2], [10], [13].

Through the practical implementation of the proposed use case, the following integration patterns between knowledge representation techniques were identified:

- **Semantic networks and Production rules:** Semantic networks represent static knowledge and the relationships between entities such as players, betting events, odds, risk, and responsible gambling measures [20], [21]. Production rules use this knowledge to make dynamic decisions based on risk thresholds, bonus conditions, and regulatory requirements [10], [24].
- **Frames and Production rules:** Frames organize detailed and structured information about individual entities, especially user accounts and betting states [1], [3], [4]. Production rules operate on this structured data to control system logic and decision-making [10], [24].
- **Scripts and Semantic networks:** Scripts represent the sequence of events during a betting session, while semantic networks capture the relationships between the entities involved in those events [12], [20], [21], [23].
- **Overall integration:** This combination creates a clear separation of roles: semantic networks model relationships, frames structure data, scripts describe behavior and event flow, and production rules enforce decisions. As a result, the system can achieve improved reliability, better risk management, decision transparency, and adaptability to future business or regulatory changes [1], [2], [8], [10], [13], [17].

V. DISCUSSION

The findings indicate that knowledge-based AI approaches, implemented through knowledge representation techniques, offer distinct and complementary advantages when applied to AI-enabled online gambling and betting systems [1], [2], [10], [13], [17]. Each representation technique contributes to a specific aspect of system intelligence, while also exhibiting limitations that reduce its effectiveness when used in isolation. Semantic networks capture relationships among players, bets, events, odds, and risk factors [20], [21]. Their graph-based structure improves interpretability and supports basic inference, which is important in regulated environments [10]. However, they face scalability issues in complex domains [21] and cannot adequately represent the uncertainty and probabilistic outcomes inherent in gambling systems [8]. Frames complement this by organizing entity-specific data such as user profiles and system states [1], [3], [4]. They support consistency and traceability through defined attributes and defaults [10], [13], but are less effective for modeling complex inter-entity

relationships [21] and may suffer from design and maintenance issues in large-scale systems [4]. Scripts model behavioral sequences and interaction flows within the platform [12], [23]. They enable process understanding, reasoning under incomplete information, and detection of anomalous behavior relevant to risk analysis [8], [17], but are limited by rigidity when handling highly dynamic or unforeseen user behavior. Production rules encode decision logic and policy constraints such as risk thresholds and responsible gambling measures [8], [10], [24]. While suitable for regulatory compliance, rule-based systems can become complex and difficult to maintain as system size increases [24]. Overall, the results highlight that the main strength of the proposed approach lies in integrating multiple knowledge representation techniques [1], [13]. By combining relational, structural, behavioral, and decision-making representations [20], [21], [23], [24], the framework forms a coherent and explainable AI system [2], [10], [13]. This integration improves reliability, risk management, and transparency compared to single-technique approaches [2], [8], [17], reinforcing the relevance of symbolic and hybrid AI in regulated gambling environments.

VI. CONCLUSION

This study presents a structured, use-case-based analysis of AI knowledge representation techniques in online gambling and betting systems, focusing on semantic networks, frames, scripts, and production rules. The analysis demonstrated that knowledge-based AI approaches are highly relevant to modern gambling platforms that require transparency, reliability, and regulatory compliance [1], [10], [13]. Each technique provides distinct capabilities: semantic networks support relationship reasoning, frames manage structured entity data, scripts capture interaction processes, and production rules enable interpretable decision logic [20], [21], [24]. However, their individual limitations indicate that no single technique can fully address the complexity of online gambling systems [13]. A key conclusion is that integrating multiple knowledge representation techniques offers the most effective solution. By combining relational, structural, behavioral, and procedural representations, the proposed approach supports more robust decision-making, improved risk management, and clearer enforcement of responsible gambling policies. This integrated framework enhances system explainability and control, addressing limitations of purely data-driven AI approaches [2], [8]. Future work may focus on developing and validating hybrid reasoning architectures that integrate symbolic knowledge representations with learning-based components under regulatory constraints, ensuring real-time auditability and decision traceability in operational gambling environments.

REFERENCES

[1]. J. Borrego-Díaz and J. Galán Páez, “Knowledge representation for explainable artificial intelligence: Modeling foundations from complex systems,” Complex & Intelligent Systems, vol. 8, no. 2, pp. 1579–1601, 2022.

[2]. K. Coussement, M. Z. Abedin, M. Kraus, S. Maldonado, and K. Topuz, “Explainable AI for enhanced decision-making,” Decision Support Systems, vol. 184, p. 114276, 2024.

[3]. M. E. Vidal, Y. Chudasama, H. Huang, D. Purohit, and M. Torrente, “Integrating knowledge graphs with symbolic AI: The path to interpretable hybrid AI systems in medicine,” Journal of Web Semantics, vol. 84, p. 100856, 2025.

[4]. M. R. Naqvi, L. Elmhadhbi, A. Sarkar, B. Archimede, and M. H. Karray, “Survey on ontology-based explainable AI in manufacturing,” Journal of Intelligent Manufacturing, vol. 35, no. 8, pp. 3605–3627, 2024.

[5]. G. Kostopoulos, G. Davrazos, and S. Kotsiantis, “Explainable artificial intelligence-based decision support systems: A recent review,” Electronics, vol. 13, no. 14, p. 2842, 2024.

[6]. F. Mihai, O. E. Aleca, and D. M. Iordache, “AI personalization and its influence on online gamblers’ behavior,” Behavioral Sciences, vol. 15, no. 6, p. 779, 2025.

[7]. W. S. Murch, M. French, and S. Kairouz, “Online gamblers’ preferences for performance and fairness in artificial intelligence systems for gambling harm detection,” Journal of Gambling Issues, 2025.
[8]. A. Hassanniakalager and P. W. Newall, “A machine learning perspective on responsible gambling,” Behavioral Public Policy, vol. 6, no. 2, pp. 237–260, 2022.
[9]. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys (CSUR), vol. 54, no.6, pp. 1-35. 2021
[10]. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, and F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, 2020.
[11]. P. Lopes, E. Silva, C. Braga, T. Oliveira, and L. Rosado, “XAI systems evaluation: A review of human and computer-centered methods,” Applied Sciences, vol. 12, no. 19, p. 9423, 2022.
[12]. S. Xiao, S. Jhaver, and N. Salehi, “Addressing interpersonal harm in online gaming communities: The opportunities and challenges for a restorative justice approach,” ACM Transactions on Computer-Human Interaction, vol. 30, no. 6, pp. 1–36, 2023.
[13]. M. Graziani, L. Dutkiewicz, D. Calvaresi, J.P. Amorim, K. Yordanova, and M. Vered, "A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences," Artificial Intelligence Review, vol. 56, pp. 3473–3504, 2022.
[14]. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A review of machine learning interpretability methods,” Entropy, vol. 23, no. 1, p. 18, 2020.
[15]. F. Binesh and K. Ghaharian, “Identifying risks and ethical considerations of AI in gambling: A scoping review,” International Journal of Hospitality & Tourism Administration, pp. 1–31, 2025.
[16]. V. Modi and A. Chanda, Responsible AI in Real-Money Gaming, 2025.
[17]. V. Marionneau, K. Ristolainen, and T. Roukka, “Duty of care, data science, and gambling harm: A scoping review of risk assessment models,” Computers in Human Behavior Reports, vol. 18, p. 100644, 2025.
[18]. D. Farrugia, C. Zerafa, T. Cini, B. Kuasney, and K. Livori, "A real-time prescriptive solution for explainable cyber-fraud detection within the iGaming industry," SN Computer Science, 2(3), 215, 2021.
[19]. N. Binesh, K. Ponnada, and A. Syah, “The future of the gambling industry is AI: Insights from expert interviews on human-AI collaboration, regulation, and ethics,” Journal of Gambling Studies, pp. 1–18, 2025.
[20]. J. Han, S. Sarica, F. Shi, and J. Luo, “Semantic networks for engineering design: State of the art and future directions,” Journal of Mechanical Design, vol. 144, no. 2, p. 020802, 2022.
[21]. A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. De Melo, C. Gutierrez, and A. Zimmermann, “Knowledge graphs,” ACM Computing Surveys, vol. 54, no. 4, pp. 1–37, 2021.
[22]. S. Sarica, J. Han, and J. Luo, "Design representation as semantic networks," Computers in Industry, 144, 103791, 2023.
[23]. S. R. Sahu, M. D. UH, M. R. TV, and M. P. Swetha, Artificial Intelligence: A Modern Approach. BR Publications, 2026.
[24]. Z. Zeng, Q. Cheng, and Y. Si, “Logical rule-based knowledge graph reasoning: A comprehensive survey,” Mathematics, vol. 11, no. 21, p. 4486, 2023.
[25]. R. Calegari, G. Ciatto, and A. Omicini, “On the integration of symbolic and sub-symbolic techniques for XAI: A survey,” Intelligenza Artificiale, vol. 14, no. 1, pp. 7–32, 2020.
[26]. M. Kejriwal, Domain-specific knowledge graph construction. Cham: Springer International Publishing. 2019.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Spatiotemporal Pattern Detection Of Precipitation Using Machine Learning And Remote Sensing Data

Ramón Fernando López Osorio
Facultad de Informática Culiacán, Universidad Autónoma de Sinaloa. Culiacán Rosales, Sinaloa, México
ferrlop@uas.edu.mx
ORCID:0007-0981-5539

Lidia Yadira Pérez-Aguilar
Facultad de Informática Culiacán, Universidad Autónoma de Sinaloa. Culiacán Rosales, Sinaloa, México
lidiaperez@uas.edu.mx
ORCID:0003-4467-0690

María Alejandra Quintero Morales
Facultad de Biología, Universidad Autónoma de Sinaloa. Culiacán Rosales, Sinaloa, México
alejandra.qm@uas.edu.mx
ORCID:0002-1435-4921

Evangelina Ávila-Aceves
Facultad de Ciencias Naturales y Exactas, Universidad Autónoma de Sinaloa. Culiacán Rosales, Sinaloa, México
evangelinaavila.facite@uas.edu.mx
ORCID:0003-2382-5470

Yedid Guadalupe Zambrano-Medina
Facultad de Ciencias Naturales y Exactas, Universidad Autónoma de Sinaloa. Culiacán Rosales, Sinaloa, México
Yedidzambrano@uas.edu.mx
ORCID:0001-8820-0688

Sergio Alberto Monjardin-Armenta
Facultad de Ciencias Naturales y Exactas, Universidad Autónoma de Sinaloa. Culiacán Rosales, Sinaloa, México
sa.monjardin12@info.uas.edu.mx
ORCID:0002-4890-6798

Abstract
Precipitation variability represents one of the main factors influencing water availability, environmental dynamics, and climate behavior in arid and semi-arid regions. This study aims to identify and characterize spatiotemporal precipitation patterns in Northwestern Mexico during the period 1960–2024 using remote sensing data and unsupervised Machine Learning techniques. K-Means, DBSCAN, and PCA were applied to identify climatically homogeneous precipitation regions and dominant patterns of regional climate variability. The results revealed significant spatial variability, with lower precipitation values in Baja California and Sonora and higher precipitation accumulation in Sinaloa. K-Means identified four optimal precipitation clusters (Silhouette = 0.4029), while PCA explained 94.33% of the total variability. Beyond pattern detection, the proposed framework provides a reproducible and interpretable approach for identifying climatically homogeneous regions, supporting hydroclimatic assessment, regional climate monitoring, and evidence-based water resource management in arid and semi-arid environments.

Keywords: Precipitation variability, machine learning, K-Means, DBSCAN, PCA, geospatial analysis, Northwestern Mexico**.**

I. INTRODUCTION

Precipitation is a fundamental component of the hydrological cycle. It plays a key role in water availability, ecosystem functioning, and environmental sustainability, especially in arid and semi-arid regions where water availability depends directly on rainfall events [1], [2]. In recent years, climate change has significantly altered precipitation patterns, increasing the frequency and intensity of droughts and extreme hydroclimatic events,

thereby affecting environmental stability, agricultural activities, and water resource availability [3], [4]. Advances in remote sensing have strengthened climate monitoring through high-resolution spatial and temporal data on atmospheric and hydrological variables [5]. Among these products, TerraClimate stands out as a widely used source in environmental and hydrological studies due to its global coverage and approximate 4 km spatial resolution [6], [7], enabling geospatial analyses focused on climate monitoring and regional hydrological variability assessment [8], [9], [10]. Machine Learning (ML) techniques have demonstrated strong potential for analyzing climatic and geospatial data by identifying spatial patterns and climate variability through unsupervised learning algorithms. Methods such as K-Means [11], [12], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [13], [14], [15], and Principal Component Analysis (PCA) [16], [17] facilitate automated precipitation pattern recognition and strengthen environmental and climate monitoring at regional scales [18]. Therefore, this study aims to identify and characterize spatiotemporal precipitation patterns in Northwestern Mexico using TerraClimate-derived precipitation data and unsupervised Machine Learning techniques for the period 1960–2024. Specifically, K-Means, DBSCAN, and Principal Component Analysis (PCA) were applied to identify climatically homogeneous regions, characterize dominant precipitation variability patterns, and evaluate the potential of integrating remote sensing and Machine Learning for hydroclimatic assessment and regional climate monitoring. The proposed framework contributes to a better understanding of precipitation variability and supports climate monitoring and water resource management in arid and semi-arid regions.

II. MATERIALS AND METHODS

The study area is located in Northwestern Mexico and includes the states of Baja California, Baja California Sur, Sonora, and Sinaloa (Figure 1) [19]. The region exhibits significant climatic diversity, with predominantly arid and semi-arid conditions in Baja California, Baja California Sur, and Sonora. In contrast, Sinaloa and the mountainous areas of the Sierra Madre Occidental are wetter, influenced by the North American Monsoon [20], [21]. Annual precipitation shows strong spatial heterogeneity, ranging from less than 150 mm in desert regions to more than 700 mm in mountainous areas and southern Sinaloa [22]. Additionally, high temperatures and regional hydroclimatic variability generate strong dependence on water resources, particularly for agricultural and environmental activities [23].

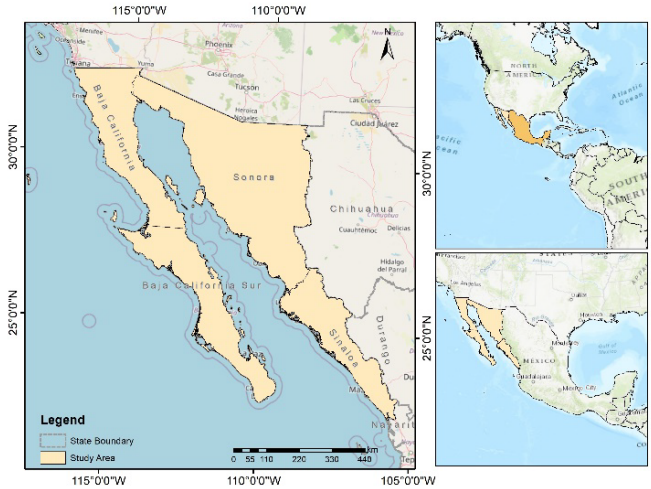


Figure 1. Location of the study area in Northwestern Mexico, including the states of Baja California, Baja California Sur, Sonora, and Sinaloa

For this research, annual accumulated precipitation derived from TerraClimate [6], [7], a widely used global climate dataset in environmental and hydrological studies [8], [9], [10], was used as the input variable for the Machine Learning analyses. TerraClimate provides climatic data at an approximate spatial resolution of 4 km for the period 1960–2024, enabling the analysis of spatiotemporal precipitation patterns in Northwestern Mexico. Precipitation was selected as a key indicator of hydroclimatic variability and water availability in arid and semi-arid regions.

III. METHODOLOGY

1. Data preprocessing

Before implementing the ML models, TerraClimate precipitation data underwent preprocessing and spatial standardization using GIS tools. All raster files were reprojected to the Mex_INEGI_Lambert_Conformal_Conic coordinate system with a uniform spatial resolution of 4644.52 m × 4644.52 m, using bilinear interpolation because precipitation data are continuous. Subsequently, raster layers were aligned to the same spatial extent, cell size, and grid structure, then clipped to the study area, ensuring spatial consistency and compatibility among the climatic datasets used in the analysis.

2. Construction of precipitation database

Subsequently, annual precipitation raster values were extracted using GIS tools to construct a geospatial database for spatiotemporal analysis and the implementation of ML models. Each record included spatial coordinates and precipitation values for the period 1960–2024, resulting in a geospatial database in which each annual precipitation layer (1960–2024) was treated as an input variable for subsequent clustering and Principal Component Analysis.

3. Spatiotemporal analysis

Spatiotemporal analysis was performed by evaluating the spatial distribution and temporal variability of precipitation in Northwestern Mexico. Initially, mean annual precipitation maps were generated to identify regional patterns and spatial contrasts among arid, semi-arid, and subhumid regions. Subsequently, interannual precipitation variations during the 1960-2024 period were analyzed to identify relatively dry and wet years and patterns associated with regional climate variability. This analysis provided a spatial basis for the subsequent implementation of clustering and dimensionality reduction techniques using ML models.

4. Machine learning models

To identify spatial precipitation patterns, unsupervised ML techniques were implemented. The models enabled the grouping of regions with similar precipitation behavior, the detection of climatic anomalies, and the analysis of spatial precipitation patterns within the study area. Additionally, PCA was applied to summarize the dominant patterns of temporal precipitation variability before interpreting the spatial structure identified by the clustering algorithms. The models were implemented in Python using Jupyter Notebook and libraries such as NumPy, Pandas, Scikit-learn, and Matplotlib for data processing, statistical analysis, spatial clustering, and visualization generation.

5. K-Means clustering

The K-Means algorithm was used to identify spatial clusters with similar precipitation patterns using unsupervised learning. This method automatically classifies climatic observations into clusters based on spatial and precipitation similarities, thereby identifying regions with homogeneous pluviometric behavior within the study area [24], [25]. Variables were normalized to reduce scale differences and prevent attributes with larger magnitudes from disproportionately influencing the clustering process [26], using equation 1, as follows:

$$x'_{ji} = \frac{x_{ji} - \boldsymbol{\mu}_j}{\boldsymbol{\sigma}_j} \tag{1}$$

where x'_{ji} represents the normalized value of the precipitation variable, x_{ji} corresponds to the original value, $\boldsymbol{\mu}_j$ represents the mean, and $\boldsymbol{\sigma}_j$ the standard deviation of the climatic observations within the study area. Subsequently, the K-Means algorithm grouped the observations by minimizing the distance between each point and the corresponding cluster centroids through an iterative spatial reassignment procedure [18] using equation 2:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left| | x_j - \boldsymbol{\mu}_j | \right|^2 \tag{2}$$

where k represents the number of clusters, C_i corresponds to the set of observations belonging to cluster i, x_j represents each climatic observation, and $\boldsymbol{\mu}_j$ corresponds to the cluster centroid.

6. DBSCAN clustering

The DBSCAN algorithm was implemented to detect homogeneous spatial regions through spatial density analysis [16], [27]. This approach enabled the identification of natural precipitation clusters and the detection of potential outliers or regions with differentiated climatic behavior. DBSCAN defines clusters based on the spatial density of observations within a given neighborhood [28], using equation 3, as follows:

$$N_\varepsilon(p) = \{q \in \mathrm{D} \mid dist\,(p,q) \leq \varepsilon\} \tag{3}$$

where $N_\varepsilon(p)$ represents the neighborhood of point p, D corresponds to the maximum search distance, and $dist\,(p,q)$represents the distance between spatial observations. The ability of DBSCAN to handle complex spatial distributions and detect anomalous observations proved useful in analyzing regional climatic patterns within the study area.

7. Principal Component Analysis

Principal Component Analysis (PCA) was applied to summarize the dominant patterns of precipitation variability across the study area by transforming the original precipitation observations into a reduced set of uncorrelated components [29], [30]. This facilitated the interpretation of the major precipitation gradients while reducing data complexity. PCA transforms the original set of variables into a new set of uncorrelated principal components using equation 4:

$$z = x * w \quad (4)$$

where x represents the original climatic data matrix, w corresponds to the eigenvector matrix, and z represents the resulting principal components. This procedure enabled the identification of the principal gradients of precipitation variability and facilitated the interpretation of regional climatic patterns represented by the first principal components.

8. Model evaluation

The evaluation of the models was performed through spatial cluster interpretation, comparative analysis, and visual validation of the obtained patterns. The results generated by K-Means, DBSCAN, and PCA were compared to identify similarities and differences in the spatial distribution of precipitation patterns. Additionally, the clusters were analyzed through cartographic representation and spatial evaluation of homogeneous regions, thereby validating the geographic and climatic coherence of the detected patterns. Furthermore, the performance of the K-Means algorithm was evaluated using the Silhouette index for different cluster configurations. At the same time, PCA was validated through the cumulative explained variance of the principal components.

IV. RESULTS

1. Spatiotemporal precipitation variability

The spatiotemporal analysis of precipitation revealed marked climatic variability in Northwestern Mexico during the 1960–2024 period (Figure 2). Spatial results showed lower precipitation in Baja California and Baja California Sur, associated with arid conditions, whereas Sinaloa and southern Sonora exhibited higher precipitation, influenced by the North American Monsoon (Figure 2a). Additionally, temporal analysis revealed interannual precipitation fluctuations with alternating wet and dry periods, reflecting high regional climatic variability and differences in precipitation stability within the study area (Fig. 2b).

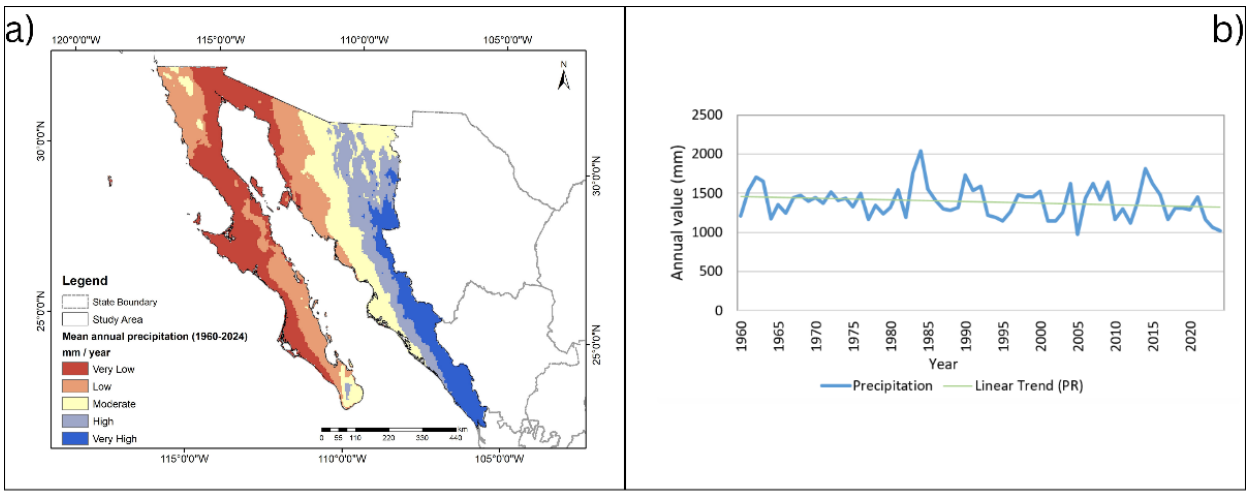


Figure 2. a) Spatial distribution of mean annual precipitation within the study area; b) temporal variability of annual precipitation during the period 1960–2024

2. K-Means clustering results

The K-Means results identified four spatial precipitation clusters in Northwestern Mexico (Figure 3). The optimal model configuration reached a Silhouette index of 0.4029, indicating acceptable cluster separation despite the gradual climatic transitions that characterize precipitation patterns across the study area. Arid regions of Baja California and Sonora were characterized by low precipitation, whereas Sinaloa and southern Sonora exhibited higher precipitation totals. Additionally, transition zones with intermediate precipitation conditions were identified, demonstrating K-Means' ability to delineate climatically homogeneous regions.

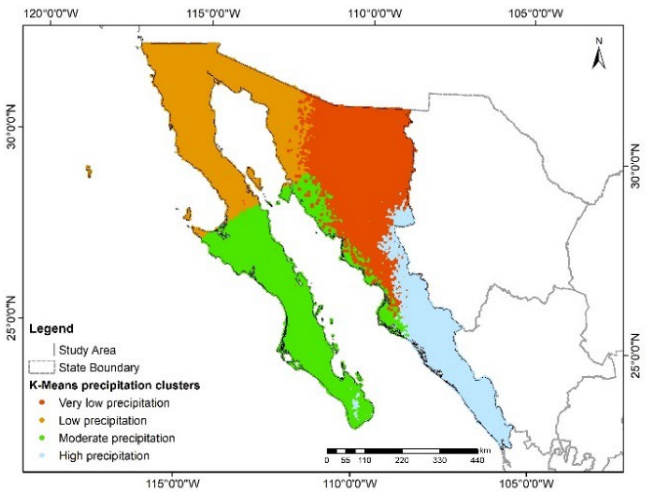


Figure 3. Spatial precipitation clusters identified using K-Means clustering in Northwestern Mexico, showing regions with very low, low, moderate, and high precipitation patterns

3. DBSCAN clustering results

The DBSCAN algorithm identified spatial density patterns associated with the regional distribution of precipitation (Figure 4). The results showed a dominant precipitation regime accompanied by spatial anomalies and secondary precipitation clusters. The anomalies were mainly located in regions with extreme precipitation behavior and high climatic heterogeneity. The ability of DBSCAN to detect irregular spatial structures complemented the results obtained with K-Means and enabled the representation of complex climatic patterns within Northwestern Mexico.

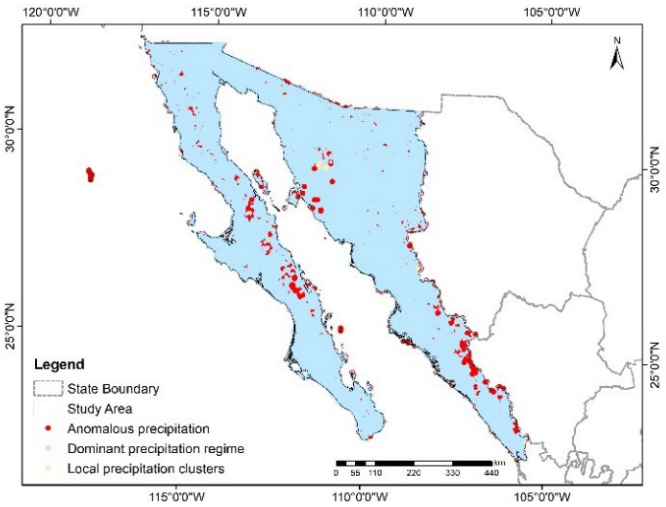


Figure 4. Spatial precipitation patterns and anomalies identified using DBSCAN clustering in Northwestern Mexico

4. Principal Component Analysis (PCA)

PCA reduced the dimensionality of the precipitation dataset and highlighted the dominant patterns of regional climatic variability (Figure 5). The first two principal components explained 94.33% of the total variance, indicating that most of the variability in precipitation can be represented by two dominant climatic gradients. These gradients differentiate the arid conditions of Baja California and Sonora from the wetter regions of Sinaloa, reflecting the influence of regional climatic controls such as topography and the North American Monsoon. Consequently, PCA simplified the interpretation of precipitation variability while preserving the main climatic information contained in the dataset.

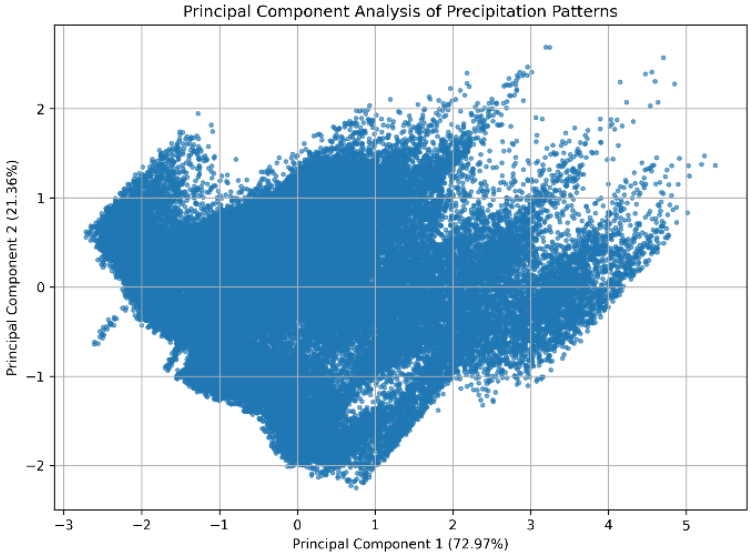


Figure 5. Principal Component Analysis (PCA) of precipitation variability in Northwestern Mexico. The first two principal components explained 94.33% of the total variance.

5. Comparison of clustering methods

The comparison between K-Means, DBSCAN, and Principal Component Analysis (PCA) revealed important differences in the spatial and statistical representation of precipitation patterns in Northwestern Mexico. K-Means facilitated the delimitation of regions with homogeneous precipitation behavior, while DBSCAN detected irregular spatial structures and regions with atypical climatic behavior. PCA complemented the analysis by reducing dimensionality and capturing the main spatial precipitation gradients, explaining 94.33% of the total variability in the data with the first two principal components. Overall, the complementary use of K-Means, DBSCAN, and PCA provided a more comprehensive characterization of precipitation variability than any individual method. While K-Means identified climatically homogeneous regions, DBSCAN detected spatial anomalies, and PCA summarized the dominant climatic gradients, they improved hydroclimatic interpretation and supported regional climate monitoring.

V. DISCUSSION

The results revealed marked spatial heterogeneity in precipitation in Northwestern Mexico, particularly between arid regions in Baja California and Baja California Sur and areas with higher precipitation accumulations in Sinaloa and southern Sonora. This spatial distribution reflects the combined influence of topographic, atmospheric, and regional climatic factors on precipitation dynamics, particularly the North American Monsoon and regional altitudinal gradients. Additionally, the observed variations are consistent with previously reported hydroclimatic patterns for arid and semi-arid regions of Northwestern Mexico, where water availability strongly depends on interannual precipitation variability and extreme climatic events associated with regional atmospheric processes.

Furthermore, the complementary application of K-Means, DBSCAN, and PCA provided a more comprehensive interpretation of precipitation variability than any individual method. K-Means identified climatically homogeneous regions, DBSCAN detected spatial anomalies associated with heterogeneous precipitation behavior, and PCA summarized the dominant climatic gradients, explaining 94.33% of the total variance. These findings demonstrate that integrating geospatial analysis and Machine Learning facilitates the identification of regional hydroclimatic patterns and provides a useful framework for climate monitoring and water resource assessment in arid and semi-arid regions.

VI. CONCLUSIONS

The spatiotemporal analysis of precipitation revealed high climatic variability in Northwestern Mexico during 1960–2024, with lower precipitation in Baja California and Sonora and higher precipitation in Sinaloa. ML techniques proved effective for identifying homogeneous spatial patterns, climatic anomalies, and dominant precipitation gradients. The integration of remote sensing, geospatial analysis, and Machine Learning provides an interpretable framework for hydroclimatic assessment, supporting regional climate monitoring, water resource management, and climate adaptation in arid and semi-arid regions.
Beyond the study area, this approach can be transferred to other arid and semi-arid regions where precipitation variability strongly influences water availability and environmental vulnerability. By identifying homogeneous precipitation regions and anomalous climatic patterns, the proposed framework can support the design of

regional monitoring strategies, drought-risk assessment, and evidence-based decision-making for water resource planning.

ACKNOWLEDGEMENT

The authors would like to thank the Facultad de Informática Culiacán at the Universidad Autónoma de Sinaloa and the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) for their support in strengthening scientific research and technological development.

REFERENCES

[1] W. Li *et al.*, 'An Assessment of the Hydrological Trends Using Synergistic Approaches of Remote Sensing and Model Evaluations over Global Arid and Semi-Arid Regions', *Remote Sens.*, vol. 12, no. 23, p. 3973, Jan. 2020, doi: 10.3390/rs12233973.

[2] J. M. Grünzweig *et al.*, 'Dryland mechanisms could widely control ecosystem functioning in a drier and warmer world', *Nat. Ecol. Evol.*, vol. 6, no. 8, pp. 1064–1076, Aug. 2022, doi: 10.1038/s41559-022-01779-y.

[3] S. Mukherjee, A. Mishra, and K. E. Trenberth, 'Climate Change and Drought: a Perspective on Drought Indices', *Curr. Clim. Change Rep.*, vol. 4, no. 2, pp. 145–163, Jun. 2018, doi: 10.1007/s40641-018-0098-x.

[4] B. I. Cook, J. S. Mankin, and K. J. Anchukaitis, 'Climate Change and Drought: From Past to Future', *Curr. Clim. Change Rep.*, vol. 4, no. 2, pp. 164–179, Jun. 2018, doi: 10.1007/s40641-018-0093-2.

[5] J. Zeng, D. Long, Y. Zhang, D. Ryu, J.-P. Wigneron, and Q. Huang, 'Emerging remote sensing techniques for hydrological applications', *Remote Sens. Environ.*, vol. 332, p. 115060, Jan. 2026, doi: 10.1016/j.rse.2025.115060.

[6] TerraClimate, 'TerraClimate', Climatology Lab. Accessed: Oct. 03, 2024. [Online]. Available: https://www.climatologylab.org/terraclimate.html

[7] J. T. Abatzoglou, S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch, 'TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015', *Sci. Data*, vol. 5, no. 1, Art. no. 1, Jan. 2018, doi: 10.1038/sdata. 2017.191.

[8] L. Y. Perez-Aguilar, W. Plata-Rocha, S. A. Monjardin-Armenta, and C. Franco-Ochoa, 'Aridity Analysis Using a Prospective Geospatial Simulation Model in This Mid-Century for the Northwest Region of Mexico', *Sustainability*, vol. 14, no. 22, Art. n°. 22. Jan. 2022, doi: 10.3390/su142215223.

[9] L. Y. Perez-Aguilar, W. Plata-Rocha, S. A. Monjardin-Armenta, and R. F. López-Osorio, 'Implementation of a web-based system for monitoring and simulation of arid zones in northwestern Mexico. Region of Mexico, *Ecol. Model*, vol. 501, p. 111024, Feb. 2025, doi: 10.1016/j.ecolmodel.2025.111024.

[10] L. P. Osorio, J.-F. Mas, F. Guerra, and M. Maass, 'Analysis and modeling of deforestation processes: a case study in the Coyuquilla River Basin, Guerrero, Mexico Analysis and modeling of deforestation processes: a case study in the Coyuquilla River Basin', *Investig. Geogr.*, vol. 88, pp. 60–74, 2015, doi: 10.14350/rig. 43853.

[11] Q. Zhao, Y. Zhu, D. Wan, Y. Yu, and Y. Lu, 'Similarity Analysis of Small- and Medium-Sized Watersheds Based on Clustering Ensemble Model', *Water*, vol. 12, no. 1, p. 69, Jan. 2020, doi: 10.3390/w12010069.

[12] S. Tsegaye, T. M. Missimer, J.-Y. Kim, and J. Hock, 'A Clustered, Decentralized Approach to Urban Water Management', *Water*, vol. 12, no. 1, p. 185, Jan. 2020, doi: 10.3390/w12010185.

[13] A. C. Comrie and E. C. Glenn, 'Principal components-based regionalization of precipitation regimes across the southwest United States and northern Mexico, with an application to monsoon precipitation variability', *Clim. Res.*, vol. 10, no. 3, pp. 201–215, Dec. 1999, doi: 10.3354/cr010201.

[14] 'Principal Component (EOF) Analysis', in *International Geophysics*, vol. 100, Academic Press, 2011, pp. 519–562. doi: 10.1016/B978-0-12-385022-5.00012-9.

[15] K. Gallacher, C. Miller, E. M. Scott, R. Willows, L. Pope, and J. Douglass, 'Flow-directed PCA for monitoring networks', *Environmetrics*, vol. 28, no. 2, p. e2434, Mar. 2017, doi: 10.1002/env.2434.

[16] J. Zhang, M. Zhang, Y. Yu, and R. Yu, 'An innovative method integrating run theory and DBSCAN for complete three-dimensional drought structures', *Sci. Total Environ.*, vol. 926, p. 171901, May 2024, doi: 10.1016/j.scitotenv.2024.171901.

[17] T. Wang, C. Ren, Y. Luo, and J. Tian, 'NS-DBSCAN: A Density-Based Clustering Algorithm in Network Space', *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 5, p. 218, May 2019, doi: 10.3390/ijgi8050218.

[18] N. Bhatia, J. M. Sojan, S. Simonovic, and R. Srivastav, 'Role of Cluster Validity Indices in Delineation of Precipitation Regions', *Water*, vol. 12, no. 5, p. 1372, May 2020, doi: 10.3390/w12051372.

[19] INEGI, Instituto Nacional de Estadística y Geografía (INEGI)'. Accessed: Apr. 26, 2024. [Online]. Available: https://www.inegi.org.mx/

[20] L. Y. Perez-Aguilar, W. Plata-Rocha, S. A. Monjardin-Armenta, C. Franco-Ochoa, and Y. G. Zambrano-Medina, 'The Identification and Classification of Arid Zones through Multicriteria Evaluation and Geographic Information Systems—Case Study: Arid Regions of Northwest Mexico', *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 11, p. 720, Nov. 2021, doi: 10.3390/ijgi10110720.

[21] Conabio, 'Portal de Información Geográfica - CONABIO'.

[22] CONABIO, 'Sistema Nacional de Información sobre Biodiversidad-CONABIO', Sistema Nacional de Información sobre Biodiversidad. Accessed: Mar. 13, 2026. [Online]. Available: http://www.conabio.gob.mx/informacion/gis/

[23] INEGI, 'Espacio y datos de México - INEGI', Espacio y datos de México. Accessed: Mar. 13, 2026. [Online]. Available: https://www.inegi.org.mx/app/mapa/espacioydatos/default.aspx

[24] J. MacQueen, 'Some methods for classification and analysis of multivariate observations', in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5.1, University of California Press, 1967, pp. 281–298. Accessed: May 28, 2026. [Online]. Available: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

[25] J. A. Hartigan and M. A. Wong, 'Algorithm AS 136: A K-Means Clustering Algorithm', *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830.

[26] P. Satyanarayana and V. V. Srinivas, 'Regionalization of precipitation in data sparse areas using large scale atmospheric variables – A fuzzy clustering approach', *J. Hydrol.*, vol. 405, no. 3, pp. 462–473, Aug. 2011, doi: 10.1016/j.jhydrol.2011.05.044.

[27] M. Usman, I. S. Sitanggang, and L. Syaufina, 'Hotspot Distribution Analyses Based on Peat Characteristics Using Density-based Spatial Clustering', *Procedia Environ. Sci.*, vol. 24, pp. 132–140, Jan. 2015, doi: 10.1016/j.proenv.2015.03.018.

[28] D. Birant and A. Kut, 'ST-DBSCAN: An algorithm for clustering spatial–temporal data', *Data Knowl. Eng.*, vol. 60, no. 1, pp. 208–221, Jan. 2007, doi: 10.1016/j.datak.2006.01.013.

[29] N. R. Clark and A. Ma'ayan, 'Introduction to Statistical Methods to Analyze Large Data Sets: Principal Components Analysis', *Sci. Signal, vol.* 4, no. 190, p. tr3, Sep. 2011, doi: 10.1126/scisignal. 2001967.

[30] C. V. Singh, 'Pattern characteristics of Indian monsoon rainfall using principal component analysis (PCA)', *Atmospheric Res.*, vol. 79, no. 3, pp. 317–326, Mar. 2006, doi: 10.1016/j.atmosres.2005.05.006.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Beyond Aggregate Metrics: Population-Weighted CVaR Auditing for Hidden Performance Failures Along Continuous Attributes

Irakli Koberidze
Department of Computer Science
Tbilisi State University
Tbilisi, Georgia
Irakli.koberidze@tsu.ge
ORCID:0009-0005-7859-5096

Magda Tsintsadze
Department of Computer Science
Tbilisi State University
Tbilisi, Georgia
Magda.tsintsadze@tsu.ge
ORCID:0000-0001-9316-3295

Abstract
Aggregate metrics like F1 obscure localized degradation on subpopulations defined by continuous attributes (e.g., loan size, age, working hours). We propose a post-hoc pipeline: Nadaraya-Watson kernel smoothing estimates local F1 along the attribute; each individual receives a local estimate; population-weighted CVaR at the 5th percentile summarizes the tail. The difference between aggregate F1 and tail CVaR defines an overconfidence gap, a continuous signal quantifying how much the headline metric overstates the experience of the worst-served population. Applied to three established benchmark datasets (German Credit, COMPAS, Adult Income), we observe gaps ranging from 0.06 to 0.185 F1 points. The gap ranking is preserved across p = 1–15%, and the method is intended as a diagnostic instrument: it surfaces where aggregate performance is misleading, not why.

Keywords: Bias, Accountability, and Transparency in AI Systems, Ethical AI and Algorithmic Fairness, Digital Inequality and Inclusion

I. INTRODUCTION

Reported model performance is almost always an average. When a deployment population is heterogeneous along a continuous attribute (the size of a credit request, the age of a defendant, the number of hours worked per week), an aggregate metric can register as acceptable even while significant fractions of that population are systematically worse served. Conventional fairness audits address part of this problem by comparing performance across discrete demographic groups. Still, discretization introduces boundary arbitrariness, conflates within-group variation, and provides no natural representation of attributes that do not correspond to protected categories in legislation. We propose a pipeline that operates entirely in continuous attribute space. A Nadaraya-Watson kernel smoother [1],[2] estimates local F1 at every point along the audit attribute; Conditional Value at Risk (CVaR) [3], borrowed from financial risk management, then summarizes the tail of this distribution in a population-weighted fashion, so a narrow region with catastrophic performance but few inhabitants registers less severely than a broad region with moderate degradation affecting many. The resulting overconfidence gap, the scalar difference between aggregate F1 and tail CVaR F1, answers a specific question: by how much does the headline metric overstate performance for the worst-served population segment (defined by the 5th percentile of local F1)? The gap is a continuous signal with no binary pass/fail threshold, leaving judgments of significance to domain stakeholders.

The contributions of this paper are as follows:

- We introduce a post-hoc auditing pipeline combining Nadaraya-Watson kernel smoothing with population-weighted CVaR to measure localized F1 along a continuous attribute, requiring no modification to the training procedure or model architecture.
- We define the *overconfidence gap* as an interpretable scalar diagnostic: it surfaces where aggregate performance is misleading, but identifying why performance degrades and deciding what to do about it requires further investigation outside this paper's scope.
- We validate the pipeline on three widely used fairness benchmarks (German Credit, COMPAS, and Adult Income), observing gaps between 0.060 and 0.185 F1 points, and provide a detailed analysis of what the gap does and does not reveal about model behavior.
- We discuss the relationship between continuous-attribute CVaR auditing and existing discrete group-fairness criteria, positioning the two approaches as complementary tools in a complete audit workflow.

II. RELATED WORK

Fairness research has largely developed around categorical notions of group membership, with most existing criteria requiring predefined, discrete groups as input. This paper occupies a complementary niche: post hoc auditing of a continuous attribute, without modifying the training procedure or requiring group labels. The survey below maps the surrounding literature and clarifies where the proposed contribution fits within each strand. Group fairness metrics. Equalized odds [4], demographic parity [5], and calibration-based criteria [6] operate on discrete groups. They are complementary to, rather than superseded by, continuous-attribute auditing. Barocas et al. [5] provide a comprehensive treatment of these criteria and their inherent tensions, noting that satisfying multiple fairness definitions simultaneously is often impossible, and that the choice of criterion should be grounded in the deployment context—training-time fairness objectives. CVaR has been used as a training objective in distributionally robust optimization (DRO) to learn classifiers that are robust to worst-case data shifts [7], [8]. Hashimoto et al. [9] apply a CVaR-like worst-group objective during training to achieve fairness without demographic labels; Sagawa et al. [10] show that explicit regularization is critical for worst-case generalization under known group structure. Continuous attributes have also been addressed during training and in post-processing stages: Agarwal et al. [11] reduce fairness constraints to cost-sensitive learning, Chzhen et al. [12] derive optimal post-processing strategies for fair regression via Wasserstein barycenters, and kernel-based approaches to continuous fairness are explored in [13]. Our contribution differs from all of these in kind: CVaR is applied post hoc as an auditing lens on an already-trained model, without modifying the training procedure or applying post-processing corrections.

Slice discovery and subgroup analysis. Slice Finder [14] and related tools automate the search for underperforming subgroups defined by conjunctions of feature predicates. Our approach is complementary: it requires specifying the audit attribute but provides a smooth, population-weighted characterization rather than a combinatorial search. D'Amour et al. [15] demonstrate that underspecification in the training pipeline can produce models that appear equivalent on aggregate metrics while diverging substantially in subgroup behavior, providing further motivation for tools that probe beyond aggregate performance.

Error analysis and model cards. Mitchell et al. [16] introduce model cards as a structured format for reporting disaggregated performance. The overconfidence gap is a natural quantity to include in such documentation, providing a concise summary of distributional performance risk for continuous covariates.

III. METHOD

Figure 1 shows the two-path structure of the pipeline. The left path is the status quo: a single aggregate F1 is computed from out-of-fold predictions and reported as the headline metric. The right path is the audit contribution: a Nadaraya-Watson kernel smoother estimates local F1 at every point along the continuous audit attribute, a reliability filter removes checkpoints with insufficient effective sample size, and each individual is assigned the local F1 of their nearest reliable checkpoint. Population-weighted CVaR at the 5th percentile then summarizes the tail of this individual-level distribution. The scalar difference between the two paths, the overconfidence gap Δ, is the primary diagnostic output. The subsections below formalize each stage.

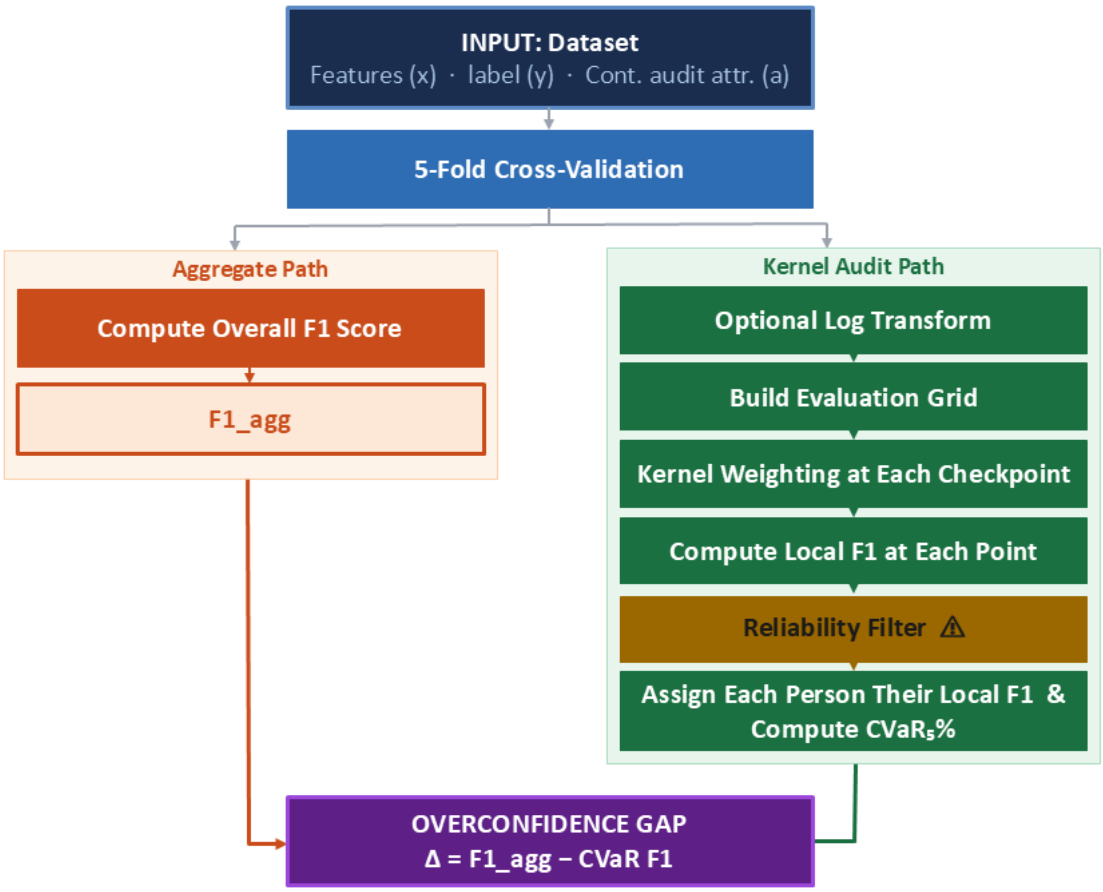


Figure 1. Diagram of the pipeline

1. Problem Setup

,D = {x_i, y_i, a}$\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^{n}$ be a dataset where $x_i \in \mathbb{R}^d$ are features, $y_i \in \{0,1\}$ is a binary label, and $a_i \in \mathbb{R}$ is the continuous *audit attribute*. Classifier predictions $\hat{y}_i \in \{0,1\}$ are obtained via 5-fold stratified cross-validation to avoid train-test leakage; aggregate F1 is computed over these out-of-fold predictions.

We optionally apply a monotone transform $t_i = \varphi(a_i)$ $(e.g., \varphi = log\ for\ right-skewed\ attributes)$ $before\ smoothing, and\ map\ the\ results\ back\ via\ the\ inverse\ \varphi\^\{-1\}\ for\ display$.

2. Local F1 via Kernel Smoothing

calculatevaluation points $\{c_1, \dots, c_M\}$ spanning [min(t), max(t)]. At each center c_j, assign soft weights

$$w_i^{(j)} = \exp\left(-\frac{1}{2}\left(\frac{t_i - c_j}{h}\right)^2\right) \quad (1)$$

where $h = 1.06\hat{\sigma}n^{-1/5}$ is the Silverman bandwidth [17].
Define weighted confusion-matrix counts

$$\widetilde{\mathrm{TP}}_j = \sum_i w_i^{(j)} \mathbf{1}[y_i = 1]\mathbf{1}[\hat{y}_i = 1] \quad (2)$$

$$\widetilde{\mathrm{FP}}_j = \sum_i w_i^{(j)} \mathbf{1}[y_i = 0]\mathbf{1}[\hat{y}_i = 1] \quad (3)$$

and analogously for $\widetilde{\mathrm{FN}}_j$, $\widetilde{\mathrm{TN}}_j$. Local F1 at c_j is then

$$\mathrm{F1_j} = \frac{2\widetilde{\mathrm{TP}}_j}{2\widetilde{\mathrm{TP}}_j + \widetilde{\mathrm{FP}}_j + \widetilde{\mathrm{FN}}_j} \quad (4)$$

Grid points are marked *unreliable* when the Kish effective sample size for weighted data [18] among positive-class observations falls below a threshold $n_{\min}$, computed from a target margin of error of ± 0.20 *at* 95% confidence: $n_{\min} = \lceil (0.5 \times 1.96 \div 0.20)^2 \rceil = 25$ [19].

3. Population-Weighted CVaR

Each i is assigned the local F1 from the nearest reliable grid point:

$$\hat{f}_i = \mathrm{F1}_{j^*(i)},\ j^*(i) = \mathrm{argmin}_{j:\mathrm{reliable}} |t_i - c_j| \quad (5)$$

This yields a distribution $\{\hat{f}_i\}$ that inherits the population density of the attribute: regions where many individuals live naturally contribute more mass.

The p-th percentile Value at Risk is $\mathrm{VaR}_p = \mathrm{Percentile}_p(\{\hat{f}_i\})$. The CVaR at level p is the mean performance among individuals whose assigned F1 falls at or below VaR_p [3]:

$$\mathrm{CVaR}_p = \mathbb{E}[\hat{f}_i | \hat{f}_i \leq \mathrm{VaR_p}] \quad (6)$$

We set p = 5, corresponding to the 95th-percentile confidence level conventional in financial risk management [3], and refer to it simply as the worst-performing tail.

The overconfidence gap is

$$\Delta = \max(0, \mathrm{F1}_{agg} - \mathrm{CVaR}) \quad (7)$$

Where $F1_{agg}$ is the aggregate cross-validated F1. A gap of zero is possible when the tail and the overall population perform similarly. Because $\Delta \geq 0$ by construction, it measures the degree to which the aggregate metric flatters performance in the tail; it cannot detect cases where the aggregate understates tail performance.

IV. EXPERIMENTAL SETUP

We apply the pipeline to three public benchmark datasets commonly used in fairness research. All classifiers are trained using 5-fold stratified cross-validation; no separate held-out test set is used. Table 1 reports the resulting gaps.

German Credit [20]. A binary task predicting credit default ($y = 1$: bad credit). We audit using log-transformed credit amounts, with Logistic Regression with balanced class weights as an interpretable baseline.

COMPAS [21]. A binary task predicting two-year recidivism ($y = 1$: recidivist). We audit along defendant age, using XGBoost with inverse class-weight scaling. Race and sex are deliberately excluded from model features; the audit could equally be applied along those axes, but we focus on the continuous numeric attribute of age in this study.

Adult Income [22]. A binary task predicting income exceeding 50K annually ($y = 1$). We audit along hours worked per week, using XGBoost with inverse class-weight scaling.

V. RESULTS

1. German Credit

Figure 2 shows local F1 as a smooth function of credit amount. Notably, the tail does not fall at the extremes but in the mid-range (2 149–2 390 DM), where 26.8 % of borrowers have bad credit. This suggests a region of mixed creditworthiness rather than a simple class-imbalance effect.

Table 1. Resulting gaps.

Dataset	Agg. F1	CVaR F1	GAP Δ	Tail Attr. Range	Tail n
German Credit	0.608	0.548	0.060	2,149-2,390 DM	56
COMPAS	0.614	0.508	0.106	57-96 Years	316
Adult Income	0.710	0.525	0.185	1-27 hrs/week	3,571

2. COMPAS

Figure 3 reveals a steady decline in local F1 with defendant age. The tail comprises defendants aged 57–96 (26.9% recidivism), consistent with reduced representation and potentially weaker generalization for older individuals. (Note: the COMPAS dataset has known limitations [21],[23]; race and sex were excluded from our model features.)

3. Adult Income

Figure 4 reveals the largest gap of the three experiments. Aggregate F1 is 0.710; CVaR is 0.525; the gap is $\Delta = 0.185$.. The tail covers workers reporting 1–27 hours per week; only 6.5% of that group earns over $50K annually.

4. Sensitivity to CVaR Percentile

Across $p \in \{1, \ldots, 15\ \%\}$ the gap ranking Adult Income > COMPAS > German Credit is preserved (Figure 5). Gap magnitudes are stable for German Credit (<2 % variation), Adult Income modestly stable (<20 %), while COMPAS varies more (48 %) because the steepest F1 decline occurs above age 57, making narrower tails more extreme. In every case, aggregate F1 materially overstates tail experience.

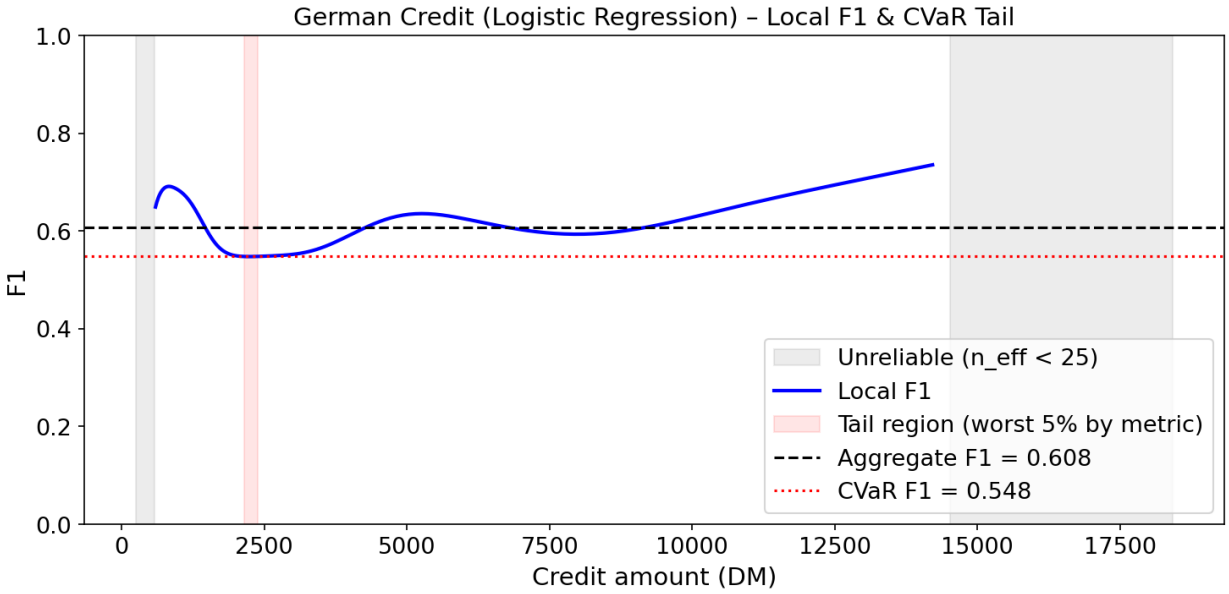


Figure 2. German Credit (Logistic Regression).

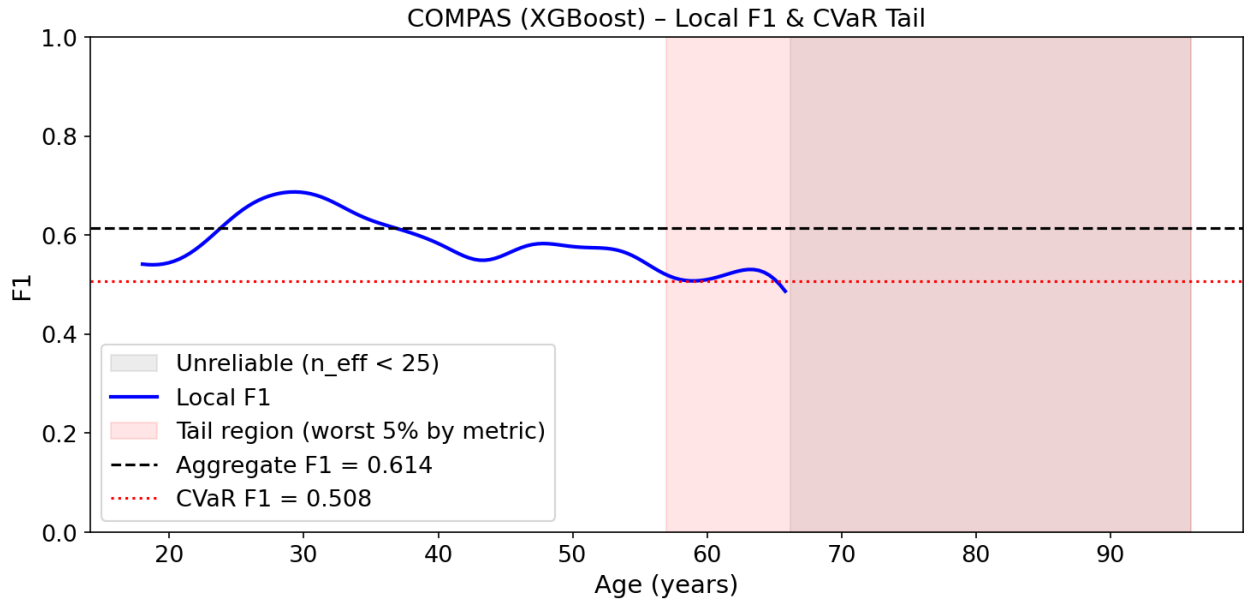


Figure 3. COMPAS (XGBoost).

VI. DISCUSSION

1. What The Gap Measures

The overconfidence gap answers a specific question: by how many F1 points the headline metric overstates performance for the worst-served population segment, weighted by how many people actually occupy that segment. This is useful for model documentation, procurement evaluation, and regulatory reporting, where a concise summary statistic is needed to characterize distributional performance risk. In German Credit, the modest gap (0.060) sits in the mid-range, reflecting mixed creditworthiness rather than a dominant predictor. COMPAS's steady decline with age (gap 0.106) warns that older defendants may see substantially worse performance than aggregate F1 suggests. Adult Income's largest gap (0.185) arises partly from structural sparsity (only 6.5% of part-time workers earn >$50K); the gap highlights a genuine disparity that aggregate F1 conceals.

2. What the Gap Does Not Measure

Cause. The pipeline identifies where performance degrades, but not why. Data sparsity, class imbalance, feature inadequacy, and distributional shift can all produce similar-looking local F1 curves. Domain expertise and targeted follow-up analysis are needed to distinguish them.
Intersectionality. The method audits one attribute at a time. It cannot detect subpopulations defined by the interaction of two or more attributes (e.g., older part-time workers), which may exhibit larger gaps than either attribute alone would suggest.

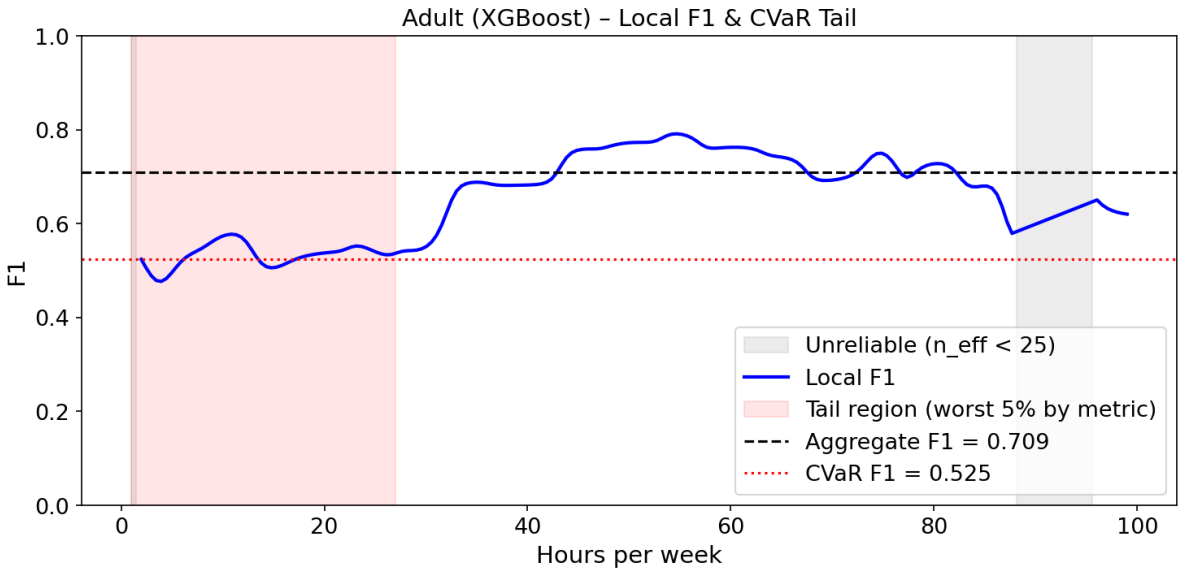


Figure 4. Adult Income (XGBoost).

Statistical uncertainty. The gap and local F1 curves are reported as point estimates; bootstrap or asymptotic confidence intervals would further increase their reliability for high-stakes decisions, and we leave this to future work.

Sensitivity to h. ***We use Silverman's rule of thumb for the bandwidth h,*** which is the default for Gaussian kernel density estimation and adapts smoothly to the data scale.

Attribute selection. The method surfaces what the analyst points it at. There is no automated procedure for identifying which attribute to audit; that choice is consequential and should be grounded in domain knowledge about where harm is likely to concentrate.

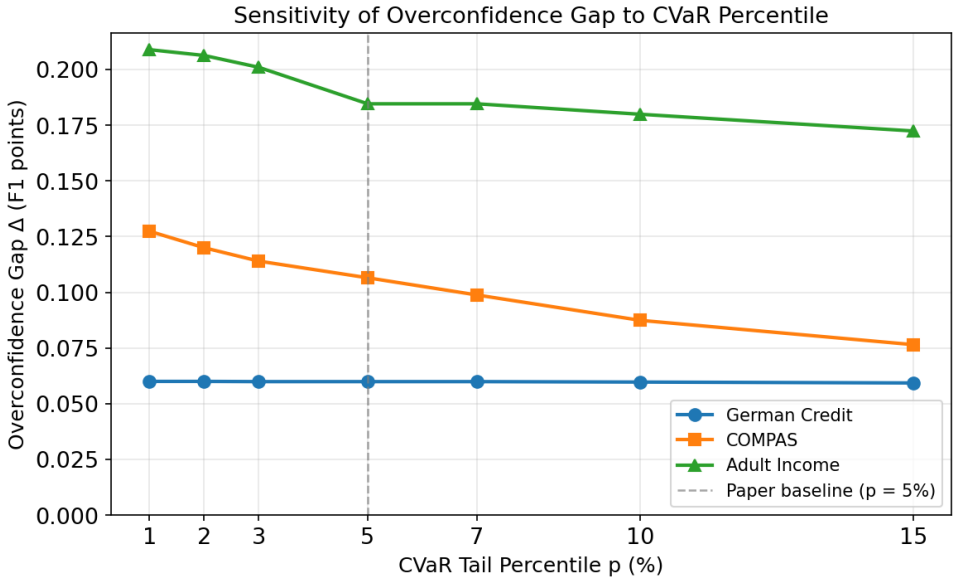


Figure 5. Sensitivity to CVaR percentile

Remediation. A smaller gap is not the same as a solved problem. The pipeline is a measurement instrument. Targeted data collection, resampling, threshold calibration by subgroup, and model retraining are among possible remediation strategies, but each introduces its own tradeoffs beyond this paper's scope.

3. Relation to Existing Fairness Criteria

The overconfidence gap is not equivalent to standard group-fairness criteria such as equalized odds or demographic parity [4], [5]. Those criteria require defining discrete groups; this method requires a continuous attribute. The two approaches are complementary: a complete audit would typically apply both, using group-fairness metrics for legally protected categories and continuous-attribute CVaR auditing for numeric covariates that correlate with harm.

VII. CONCLUSION

We have presented a post-hoc auditing pipeline that uses Nadaraya-Watson kernel smoothing and population-weighted CVaR to measure the overconfidence gap between aggregate F1 and the performance experienced by the worst-served population segment (the tail defined by the 5th percentile of local F1), identified along a continuous attribute. Applied to German Credit, COMPAS, and Adult Income, we find overconfidence gaps of 0.060, 0.106, and 0.185 (baseline p=5%); the ranking Adult Income > COMPAS > German Credit persists across $p \in \{1–15\%\}$ but gap magnitudes vary. In every case, aggregate F1 overstates the performance experienced by the worst-served population segment. We present this method as a practical instrument for audit toolkits, suited for continuous covariates known to correlate with harm, and for reporting or procurement evaluations that require a concise summary statistic.

REFERENCES

[1]. E. A. Nadaraya, “On estimating regression,” *Theory of Probability and Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.
[2]. G. S. Watson, “Smooth regression analysis,” Sankhyā: The Indian Journal of Statistics, Series A, vol. 26, no. 4, pp. 359–372, 1964.
[3]. R. T. Rockafellar and S. Uryasev, “Optimization of conditional value-at-risk,” Journal of Risk, vol. 2, no. 3, pp. 21–41, 2000.
[4]. M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 3315–3323.
[5]. S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA: MIT Press, 2023. [Online]. Available: https://fairmlbook.org.
[6]. A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” Big Data, vol. 5, no. 2, pp. 153–163, 2017.
[7]. H. Namkoong and J. C. Duchi, “Stochastic gradient methods for distributionally robust optimization with f-divergences,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 29, 2016.



[8]. D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford, “Large-scale methods for distributionally robust optimization,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.

[9]. T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in Proc. 35th Int. Conf. Machine Learning (ICML), vol. 80, 2018, pp. 1929–1938.

[10]. S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization,” in Proc. Int. Conf. Learning Representations (ICLR), 2020.

[11]. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, “A reductions approach to fair classification,” in Proc. 35th Int. Conf. Machine Learning (ICML), vol. 80, 2018, pp. 60–69.

[12]. E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, “Fair regression with Wasserstein barycenters,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 7321–7331.

[13]. A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, “Fair kernel learning,” in Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), Lecture Notes in Computer Science, vol. 10534, Springer, Cham, 2017, pp. 339–355.

[14]. Y. Chung, T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang, “Slice finder: Automated data slicing for model validation,” in Proc. 35th IEEE Int. Conf. Data Engineering (ICDE), 2019, pp. 1550–1553.

[15]. A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al., “Underspecification presents challenges for credibility in modern machine learning,” Journal of Machine Learning Research, vol. 23, no. 226, pp. 1–61, 2022.

[16]. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for model reporting,” in Proc. Conf. Fairness, Accountability, and Transparency (FAccT), Atlanta, GA, USA, 2019.

[17]. B. W. Silverman, Density Estimation for Statistics and Data Analysis. London: Chapman & Hall, 1986.

[18]. L. Kish, Survey Sampling. New York: Wiley, 1965.

[19]. W. G. Cochran, Sampling Techniques, 3rd ed. New York: Wiley, 1977.

[20]. H. Hofmann, “Statlog (German Credit Data),” UCI Machine Learning Repository, 1994. [Online]. Available: https://doi.org/10.24432/C5NC77

[21]. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “COMPAS Analysis Data and Code,” ProPublica, 2016. [Online]. Available: https://github.com/propublica/compas-analysis

[22]. B. Becker and R. Kohavi, “Adult,” UCI Machine Learning Repository, 1996. [Online]. Available: https://doi.org/10.24432/C5XW20

[23]. J. Dressel and H. Farid, “The accuracy, fairness, and limits of predicting recidivism,” Science Advances, vol. 4, no. 1, eaao5580, January 2018.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Dual-Explainer GRU–TCN for Cyber-Physical Attack Detection in IoT-Based Energy Infrastructure

Ahmet Nusret Özalp
Department of Computer Engineering
Kastamonu University
Kastamonu, Türkiye
anozalp@kastamonu.edu.tr
ORCID: 0000-0003-4882-9216

Viktoriia Hreniukh
Department of Computer Engineering
Ukrainian National Forestry University
Lviv, Ukraine
v.grenyukha@nltu.lviv.ua
ORCID: 0009-0001-8275-6762

Abstract
The integration of Internet of Things (IoT) technologies into energy production and distribution infrastructure has expanded the attack surface of industrial control systems (ICS), making cyber-physical threats a critical issue for operators of smart grids, turbine plants, and hydroelectric power plants. Traditional IT-oriented intrusion detection systems are unable to capture the nature of the physical processes characteristic of attacks in the energy sector. This paper explores the application of machine learning to detect cyber-physical attacks on HAI v21.03 – the only publicly available test dataset for energy production processes (steam turbines and pumped-storage power plants) as of 2025. We compare four model families (GRU, TCN, XGBoost, Isolation Forest) under leak-free protocols, using the proposed GRU–TCN ensemble calibrated with a single threshold, and use SHAP (GradientExplainer for GRU, TreeExplainer for XGBoost) for interpretability with dual explanations. Across ten random seeds, the ensemble achieves F1 = 0.659 and AUC-ROC = 0.891. When transferred without training on DataSense, all models drop to a level close to random (AUC 0.40–0.67): temporal models plateau at the baseline prevalence level (F1 = 0.587), while XGBoost drops to F1 = 0.000, despite having the highest AUC in the domain (0.918), which constitutes a transfer failure. SHAP at the instance level precisely matches the documented HAI attack variables (overlap@3 = 1.00). To the best of our knowledge, this is the first study to combine SHAP with dual explanations for HAI and to quantitatively transfer, without training, to an independently collected IIoT environment.

Keywords: cyber-physical attack detection, industrial control systems, HAI v21.03, GRU–TCN ensemble, SHAP explainability, zero-shot transfer, cross-domain generalization, class imbalance handling

I. INTRODUCTION

The energy sector constitutes one of the most critical components of modern society – from healthcare and transport to water supply and finance. Over the past decade, the operational technology (OT) layer of energy infrastructure, which includes supervisory control and data acquisition (SCADA) systems, distributed control systems (DCS), and programmable logic controllers (PLC), has been opened up by Internet of Things (IoT)-connected devices, cloud-edge computing architectures, and standard communication protocols. The efficiency gains are real, but so are the costs: facilities that were once physically isolated and closed now face the same threats as corporate IT systems, with the added potential for physical consequences. These are not hypothetical risks. For example, the Stuxnet (2010) campaign showed that control systems governing centrifuge processes could be compromised to cause physical damage while skipping monitoring; the 2015–2016 BlackEnergy campaigns cut power to hundreds of thousands of Ukrainian households; the 2021 Colonial Pipeline ransomware disrupted fuel supply across the eastern United States; and in the same year an intruder briefly altered chemical levels at the Oldsmar water plant in Florida. More recent activity: Industroyer2 against a Ukrainian substation in 2022, and the Volt Typhoon cluster pre-positioning inside U.S. energy and water utilities, follows the same

trajectory: attackers reach for the physical process directly, their tooling is increasingly portable across equipment, and intrusions often go unnoticed for long periods [1]. Detection, therefore, has to operate on physical sensor data and generalize to plants the model has never seen. Network-based intrusion detectors reach well above 97% accuracy on IoT traffic datasets such as CICIoT2023, TON_IoT, and X-IIoTID. Yet, those datasets describe packets and flows, not the state of a physical process. An attacker who changes a turbine reading by a fraction of a per cent per second, staying within the sensor-noise band, leaves no traffic signature at all; the anomaly lives in the temporal dynamics of the process variables [2]. Detection in energy Industrial Control Systems (ICS) thus needs models trained on process-level time series from real testbeds. The HAI (Hardware-in-the-Loop Augmented ICS Security) dataset, released by the Affiliated Institute of ETRI, is the public benchmark that directly simulates energy production processes: steam turbine and pumped-storage hydropower generation, using a physical testbed augmented by a HIL simulator. HAI v21.03 provides 79 process variables, 50 documented attack scenarios after the attack labels are removed, and a test-set anomaly rate near 2.23%. Despite its unique suitability, no prior work has applied explainable machine learning to HAI, nor tested whether models trained on it transfer to an independently collected IIoT environment.

This study is motivated by three specific research questions. RQ1: Can temporal sequence models trained on HAI provide deployable detection under a single fixed threshold, complementing higher-F1 reinforcement-learning thresholding schemes that offer neither interpretability nor cross-dataset evidence? RQ2: Do SHAP rankings point to physically meaningful variables that match the documented attack mechanisms, so that an operator can act on them? RQ3: Do the learned representations carry over to a heterogeneous IIoT testbed (DataSense) without retraining, and how large is the gap? The present paper addresses these gaps through four contributions. We build a leak-free, ten-seed comparison of four model families and a GRU–TCN ensemble on HAI v21.03; we compute dual-explainer SHAP (GradientExplainer for the GRU, TreeExplainer for XGBoost) and tie the important variables to known attacks; we quantify zero-shot and few-shot transfer to DataSense; and we give a stratified comparison of nine ICS/IIoT datasets (Table 1). Taken together, the results show that temporal models maintain useful detection under domain transfer. At the same time, tree-based detectors break under threshold transfer, and the resulting decisions remain interpretable for field operators.

II. RELATED WORK

High-quality datasets are the foundation of ICS/IIoT intrusion-detection research, and the nine key datasets are summarised in Table 1, covering SCADA water supply systems, IoT/IIoT traffic, and hybrid testbeds. BATADAL [3] simulates a water supply network with 43 variables at hourly resolution; it is widely used but does not provide network traffic and only approximates time. SWaT [4] and its extension, WADI [5], are physical water-treatment testbeds with a rich set of sensors/actuators and many implemented attacks, which remain standard ICS benchmarks even outside the energy sector. The Mississippi State gas pipeline and water tank datasets [6] are among the earliest publicly available SCADA datasets. Still, they contain known statistical issues that limit their use as a primary benchmark. As for the network component, the CICIoT2023 dataset [7] covers 33 attack types across 105 Internet of Things (IoT) devices. TON_IoT [8] adds telecommunications data and host logs for nine attack families, and X-IIoTID [9] offers Industrial Internet of Things (IIoT) traffic, independent of connection type and device, labeled according to the Cyber-Kill-Chain model; the claimed accuracy typically exceeds 97%. Still, none of these datasets encodes the physical consequences of an attack. Working with these datasets remains instructive: the LSTM-autoencoder-based threat detection model developed by Yazdinejad et al. [10] and the manifold-based adversarial analysis by Al-Fawa'reh et al. [11] illustrate both the promise of deep models for IoT data and their vulnerability to distributional shifts.

HAI [12] is the only publicly available dataset specifically created for ICS, compiled on a real-world test platform (GE turbine, Emerson boiler, FESTO water treatment system) using a HIL simulator; it contains 79 features and has a 2.23% anomaly rate over 19 days. The best published results are based on threshold adaptation: Yang et al. [13] achieve F1 = 0.994 using agent-based dynamic threshold setting on a DGN, while Homaei et al. [14] report an F1 score of 0.923 using a Causal Digital Twin, and a comparative study by Kim et al. [15] evaluates RANSynCoder at InterFusion at F1 = 82.9%. It does not apply explainable ML to HAI and does not report transferability across datasets. In DataSense [16], the most recent dataset considered, the hierarchical feature selection with the random forest algorithm achieves an F1 score of 0.97 on sensor and network data streams from an IIoT testbed containing 40 devices for 50 attacks. This is the only test set that combines synchronized data from sensors and the network and has had only one prior evaluation, making it a natural target for transferring a model trained on HAI.

Table 1. Comparative Analysis of ICS/IIoT Datasets for Attack Detection

Dataset	Samples	Feat.	Attacks	Attack %	Data type	Energy suitability
Gas Pipeline	274,627	16	~10	21.9%	Network (Modbus)	Low — no physical process sim; pattern artifacts
BATADAL	~175,000	43	14	4.7%	Physical (SCADA)	Low — water domain; hourly resolution
SWaT	449,919	51	36	12.0%	Physical + network	Medium — not energy domain
WADI	1,048,571	127	15	6.0%	Physical (SCADA/PLC)	Low-Med — stealthy attacks
HAI	1,323,608	79	50	2.2%	Physical (DCS/HIL)	Very High — only energy ICS
CICIoT2023	46,686,579	46	33	27.3%	Network (IoT)	Low — no physical part
TON_IoT	2,223,921	44	9	96.4%	Network + OS + telemetry	Low — no CP consequence
X-IIoTID	820,834	68	~30	48.7%	Network + host + system	Medium — heterogeneous IIoT
DataSense	>1.82B pkts	80+	50	Balanced	Network + sensor	High — synced; energy incl.

III. METHODOLOGY

1. System Architecture

The pipeline has four stages: ingestion and preprocessing, sliding-window feature extraction, model training with threshold calibration, and post-hoc SHAP explanation. Raw HAI time series are normalized and split into fixed-length windows; the three process-level attack-label columns (attack_P1–P3) are dropped before feature construction to prevent label leakage, leaving 79 inputs. Each window feeds four model families: the GRU and



TCN consume it as a sequence; XGBoost receives per-window statistics (mean, standard deviation, min, max, range); and the Isolation Forest operates only on flattened normal-class windows. The proposed model averages the GRU and TCN sigmoid outputs before thresholding. The decision threshold is chosen on the validation split using a 181-point grid over [0.05, 0.95], then applied unchanged to both the HAI test set and DataSense.

2. Data Preprocessing

HAI v21.03 comes from a HIL-augmented testbed that emulates a combined steam-turbine and pumped-storage process: a boiler loop (P1), a turbine with vibration monitoring (P2), a water-treatment/pumped-storage stage (P3), and a HIL simulator (P4) that ties the plant to a micro-grid model. The 79 variables cover tank levels, flow rates, pressures, temperatures, and actuator commands. All channels are min–max scaled on the training partition; the few missing values (< 0.3%) are forward-filled and then back-filled. We perform segmentation with a 30-step window (30 s at 1 Hz) and a stride of 1, which is sufficient to capture the gradual multi-step attacks documented for HAI. A window is labeled malicious if any of its steps is, and any positive rule that maximizes sensitivity but slightly inflates positive counts at attack boundaries; we keep it for comparability with prior HAI work and revisit it as a validity threat in Section VI. To avoid leakage, we split the official test files by time rather than by row. The first 70% (281,403 rows) form a development pool, and the last 30% (120,602 rows) are held out and never seen during training. The pool is divided into 20 contiguous temporal blocks; the five whose attack rate is closest to the pool mean form the validation set (70,350 rows, 2.29% attack, matching the 2.34% test rate), and the remaining blocks are added to the 921,603 normal training rows, giving 1,132,656 training rows in total. For DataSense, a separate scaler is fitted, the feature spaces are disjoint, and the HAI-trained weights and threshold are applied directly under zero-shot transfer.

3. Model Architectures

The GRU [17] stacks two layers (hidden sizes 128 and 64, dropout 0.3) over a sigmoid output (117,569 parameters). The TCN [18] uses four dilated causal blocks (dilations 1-8, 64 filters, kernel 3) with residual connections (107,905 parameters). The proposed model is a late-fusion GRU–TCN ensemble that averages both sigmoid outputs before the validation-calibrated threshold. XGBoost [19] runs 500 trees (max depth 6, subsample 0.8, colsample 0.8) with scale_pos_weight set to the window-level negative/positive ratio of 216. The Isolation Forest [20] is trained on flattened normal windows with contamination set to the training attack rate (floored at 0.005). All temporal models use the Adam optimizer (learning rate = 5×10^{-4}), Focal Loss ($\gamma = 2.0$, pos_weight = 216), and early stopping based on validation AUC-ROC (patience = 20, max = 150 epochs). SHAP TreeExplainer is applied to XGBoost; for the GRU, GradientExplainer [21] is applied with 200 normal background windows over 300 attack windows. Training ran on a single NVIDIA A100 (CUDA 12.8, 85 GB RAM); the two temporal networks together took about 8 hours across 10 seeds, with peak GPU memory below 0.2 GB.

4. Evaluation Protocol

We report the F1 score (primary metric), precision, recall, accuracy, and AUC-ROC. The F1 score is the primary metric because the test set is significantly imbalanced (2.34% attacks). AUC-ROC complements it as a threshold-independent ranking metric, which may differ from F1 in the presence of rare events. All values are means ± standard deviation across ten initial values {42, 123, 456, 789, 2024, 7, 1337, 2718, 31415, 9001}, which vary the weight initialization and batch order, with a fixed data distribution, so that the variance reflects only optimization noise. Paired differences are tested using the Wilcoxon test for F1 at each initial value ($\alpha = 0.05$); the smallest one-sided p-value at n = 10 is 0.002. The 95% confidence intervals within the domain are normal

approximation intervals: mean ± 1.96 × std/√n. The results across datasets are point estimates from a single seed, with a bootstrap sample of 2,000 iterations over the DataSense test set, used to confirm the degeneracy of predictions for saturated models [22].

IV. RESULTS

Table 2 reports in-domain performance on the held-out HAI partition (averaged across 10 seeds). The proposed ensemble gives the best F1 (0.659) and accuracy (0.984), while XGBoost holds the best precision (0.818) and AUC-ROC (0.918). The ensemble's edge over the stronger single network is small and not significant (Wilcoxon vs GRU: W = 42.0, p = 0.080), but it stabilizes predictions: ensemble precision varies far less across seeds (std = 0.113) than the TCN's (std = 0.226), so averaging the two removes variance without adding parameters. The Isolation Forest lags badly (F1 = 0.123), confirming that the 2.34% regime is too sparse for an unsupervised baseline. The contrast between XGBoost's high AUC and modest F1 (precision 0.818, recall 0.453) reflects what each metric measures: AUC rewards correct ranking, whereas F1 depends on a single operating point, and at a 2.34% base rate, the validation search settles on a cautious threshold that protects precision at the expense of recall. The temporal models accept a lower threshold that recovers more attacks (recall 0.551 for the GRU, 0.598 for the TCN), thereby achieving higher F1 despite slightly weaker ranking, and the relative ordering is stable across seeds.

Table 2. In-Domain Detection on HAI v21.03 (Mean ± Std, 10 Seeds)

Model	F1-Score	Precision	Recall	Accuracy	AUC-ROC	Inference latency
Isolation Forest	0.123±0.027	0.394±0.088	0.073±0.016	0.972±0.001	0.805±0.012	7.488 ms
XGBoost	0.583±0.014	0.818±0.025	0.453±0.015	0.983±0.001	0.918±0.004	0.623 ms
TCN	0.590±0.121	0.654±0.226	0.598±0.076	0.973±0.021	0.869±0.022	3.424 ms
GRU	0.620±0.064	0.730±0.116	0.551±0.067	0.982±0.005	0.878±0.016	2.685 ms
GRU–TCN Ensemble (proposed)	0.659±0.063	0.770±0.113	0.588±0.078	0.984±0.004	0.891±0.013	6.109 ms

95% CI for F1 (mean ± 1.96 × std/√n): IF [0.106, 0.140]; XGBoost [0.574, 0.592]; TCN [0.515, 0.665]; GRU [0.580, 0.660]; Ensemble [0.620, 0.698]. Latency: CPU, single sample, 200-iteration mean. Table 3 reports zero-shot transfer: no retraining, fine-tuning, or threshold recalibration on DataSense. Because the two testbeds expose disjoint sets of sensors, every detector drops to near-chance ranking (AUC 0.40–0.67), so F1 must be interpreted with AUC. The ensemble, TCN, and Isolation Forest saturate at the trivial flag-everything point (recall 1.000, precision 0.416, F1 = 0.587), exactly the 41.6% baseline level of spread, not genuine detection; the GRU alone avoids saturation (F1 = 0.544), but its AUC of 0.590 is near-chance. XGBoost is the telling case: despite the best in-domain AUC, it collapses to F1 = 0.000, since the estimates of its leaves are concentrated within the domain shift, and the HAI-calibrated threshold falls outside the range of DataSense estimates; each window is classified as normal. High in-domain AUC, therefore, does not guarantee cross-domain reliability for a tree-based detector.

Table 3. Zero-Shot Transfer: Trained on HAI, Tested on DataSense

Model	F1-Score	Precision	Recall	Accuracy	AUC-ROC
Isolation Forest	0.587	0.416	1.000	0.416	0.633
XGBoost	0.000	1.000	0.000	0.584	0.673
TCN	0.587	0.416	1.000	0.416	0.398
GRU	0.544	0.401	0.844	0.412	0.590
GRU–TCN Ensemble (proposed)	0.587	0.416	1.000	0.416	0.590

Models with recall = 1.000 saturate to the 41.6% prevalence baseline; F1 = 0.587 is not a genuine detection. Bootstrap F1 95% CIs have width ≤ 0.002, confirming prediction degeneracy. SHAP identifies the variables that drive the temporal model (Table 4): the level-control valves of the P3 water-treatment and P1 boiler stages (P3_LCV01D, P1_LCV01D, P1_LCV01Z), the steam-turbine load demand (P4_HT_LD), the boiler pressure-control valve (P1_PCV01D/Z), the turbine power output (P4_HT_PO), and the feed-water pump command (P3_LCP01D). The XGBoost TreeExplainer ranking overlaps on the P1_LCV01D and P1_PCV01Z channels, so two architecturally different explainers agree that the dominant attack surface is the level- and pressure-control valves of P1 and P3. The ranking is stable with sample size: the top 8 are identical for 200-300 windows and agree on six of eight between 100 and 300, with the top five unchanged across all three runs.

Table 4. Top-8 SHAP Features for the GRU (GradientExplainer)

Rank	Feature ID	Mean |SHAP|	Physical interpretation
1	P3_LCV01D	0.0129	Position command for LCV01 valve
2	P1_LCV01D	0.0071	Position command for LCV01 valve
3	P1_LCV01Z	0.0055	Current position of LCV01 valve
4	P4_HT_LD	0.0043	Electrical load demand for the steam-turbine model
5	P1_PCV01D	0.0039	Position command for PCV01 valve
6	P1_PCV01Z	0.0039	Current position of PCV01 valve
7	P4_HT_PO	0.0032	Electrical power output of the steam-turbine model
8	P3_LCP01D	0.0032	Speed command for the feed water pump

At the instance level, SHAP over the timesteps of a representative held-out attack window (Figure 1) yields three leading contributors: P3_LCV01D, P1_B3004, P1_LCV01D—that match the documented HAI attack variables exactly: overlap@3 = 1.00, overlap@5 = 0.80, overlap@8 = 0.62, with P3_LCV01D dominating at timesteps 27–28, the documented onset. Because the recovered variables (P1_LCV01D and setpoints B3004/B3005) are identified as HAI attack targets, the explanation runs on existing SCADA telemetry without retraining, and the top channels can be monitored with tighter alarms while the rest serve as corroboration.

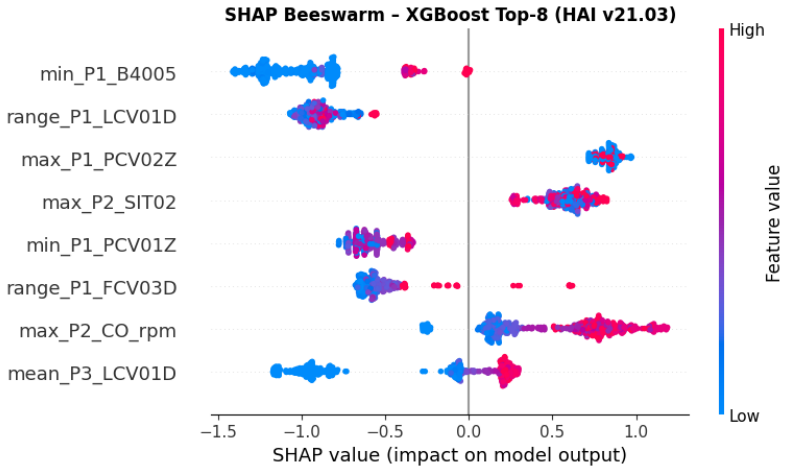


Figure 1. Instance-level SHAP over time for a representative HAI v21.03 attack window.

On HAI, dynamic-thresholding methods report higher F1 scores: ADT/DQN [12] reaches 0.994, and the causal digital twin of Homaei et al. [13] reaches 0.923, while RANSynCoder and InterFusion [14] sit at 0.829 and 0.789, respectively, but none offer interpretability or cross-dataset evidence. Our ensemble uses a single fixed threshold, which is more realistic to deploy. It trades a few points of raw in-domain F1 for dual-explainer interpretability, ten-seed confidence intervals, and the first cross-dataset transfer study on HAI.

V. DISCUSSION

The transfer results carry the practical lesson. With disjoint sensors, the near-chance AUC is mainly a representational gap that thresholding cannot fix; XGBoost's collapse to zero F1 adds a calibration failure on top, since gradient-boosted trees are not probability-calibrated, and a threshold fixed at HAI's 2.34% prevalence is meaningless at DataSense's 41.6%. This undocumented, to our knowledge, threshold-transfer failure in tree-based ICS detectors argues against selecting models by in-domain AUC. It motivates post-hoc calibration (e.g., Platt scaling or isotonic regression) before any cross-domain use. Few-shot tuning confirms the gap is representational, with the GRU rising only from 0.544 to 0.572, 0.576, and 0.588 at 10, 50, and 100 labels, respectively. The framework is therefore a complement to high-F1 dynamic-thresholding methods, trading raw F1 for interpretability (overlap@3 = 1.00), calibrated uncertainty, and an explicit map of where detection fails.

VI. LIMITATIONS

These results come with a few warnings. HAI is a single hardware-in-the-loop testbed built on specific commercial equipment (GE turbine, Emerson DCS, FESTO modules); therefore, transfer to field units with custom firmware has not been tested. The cross-dataset evaluation is deliberately zero-shot across disjoint feature spaces; it exposes the threshold-transfer failure but cannot fully separate representational from distributional gaps. Adversarial robustness (e.g., FGSM/PGD on sensor channels) was not tested, and instance-level SHAP was shown for a single representative window rather than across all 50 episodes. Hence, Table 4 is indicative rather than definitive. Deployment costs are only partially characterized: CPU latency is reported, but on-gateway energy and memory were not measured. Finally, with 10 seeds, the Wilcoxon test has limited power, so the ensemble's edge ($p = 0.080$) is a trend rather than a confirmed effect; $n \geq 30$ would be needed for confirmation.

VII. CONCLUSION

We presented an explainable machine-learning framework for cyber-physical attack detection in IoT-enabled energy infrastructure, evaluated on HAI v21.03 under leak-free, ten-seed protocols. A late-fusion GRU–TCN ensemble gave the best in-domain F1 (0.659, AUC-ROC 0.891), modestly ahead of the GRU and TCN and far above the Isolation Forest, while XGBoost held the highest AUC (0.918); the ensemble's advantage was directionally consistent (Wilcoxon p = 0.080) rather than significant. Dual-explainer SHAP converged on the level- and pressure-control valves of the P1 and P3 subsystems, with instance-level attributions matching the documented HAI attack variables exactly (overlap@3 = 1.00), the first such analysis on this benchmark. A zero-shot and few-shot study on DataSense mapped the limits of transfer. All models fell to near-chance ranking, the temporal detectors saturated to the prevalence baseline, and XGBoost collapsed under threshold transfer. Interpretable in-domain detection is achievable today, but cross-testbed deployment needs explicit domain adaptation rather than zero-shot reuse. Future work points in three directions. On the model side, attention-based and Transformer encoders, well-suited to the long-range dependencies of 1 Hz ICS data, warrant evaluation on HAI, alongside hybrid graph–temporal models that exploit sensor topology. On the transfer side, lightweight post hoc calibration (Platt scaling, isotonic regression) and adaptive thresholding should be tested as cheap remedies for the failure observed here, alongside further validation targets (SWaT, WADI, X-IIoTID) toward a multi-domain generalization matrix. Finally, embedding interpretability into the training objective rather than computing it post hoc may improve both the reliability of explanations and robustness to sensor manipulation.

REFERENCES

[1]. F. Marković and A. Kovačević, "Cyber threats and energy security: development and analysis of an incident dataset for 2022–2024," Energy Policy, vol. 208, 114913, 2026.

[2]. Özalp, Ahmet Nusret, et al. "Layer-based examination of cyber-attacks in IoT." *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, 2022.

[3]. R. Taormina, S. Galelli, N. O. Tippenhauer, E. Salomons, and A. Ostfeld, "Battle of the attack detection algorithms: disclosing cyber attacks on water distribution networks," J. Water Resour. Plann. Manage., vol. 144, no. 8, 04018048, 2018.

[4]. J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in Critical Information Infrastructures Security. Springer, 2017, pp. 88–99.

[5]. C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: a water distribution testbed for research in the design of secure cyber-physical systems," in Proc. 3rd Int. Workshop CySWATER, 2017, pp. 25–28.

[6]. T. Morris and W. Gao, "Industrial control system traffic data sets for intrusion detection research," in Int. Conf. Critical Infrastructure Protection. Springer, 2014, pp. 65–78.

[7]. E. C. P. Neto et al., "CICIoT2023: a real-time dataset and benchmark for large-scale attacks in IoT environment," Sensors, vol. 23, no. 13, 5941, 2023.

[8]. N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: network TON_IoT datasets," Sustain. Cities Soc., vol. 72, 102994, 2021.

[9]. M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, "X-IIoTID: a connectivity-agnostic and device-agnostic intrusion dataset for Industrial Internet of Things," IEEE Internet Things J., vol. 9, no. 5, pp. 3962–3977, 2022.

[10]. A. Yazdinejad et al., "An ensemble deep learning model for cyber threat hunting in Industrial Internet of Things," Digit. Commun. Netw., vol. 9, pp. 101–110, 2023.

[11]. M. Al-Fawa'reh et al., "On and off the manifold: generation and detection of adversarial attacks in IIoT networks," J. Netw. Comput. Appl., vol. 235, 104102, 2025.

[12]. H.-K. Shin, W. Lee, J.-H. Yun, and B. G. Min, "Two ICS security datasets and anomaly detection contest on the HIL-based augmented ICS testbed," in Proc. Cyber Security Experimentation and Test Workshop (CSET '21), 2021, pp. 36–40.

[13]. X. Yang, E. Howley, and M. Schukat, "ADT: time series anomaly detection for cyber-physical systems via deep reinforcement learning," Comput. Secur., vol. 141, 103825, 2024.

[14]. M. Homaei, M. Tarif, P. G. Rodriguez, A. Caro, and M. Avila, "Causal digital twins for cyber-physical security in water systems," Mach. Learn. Appl., vol. 23, 100824, 2025.

[15]. B. Kim, M. A. Alawami, E. Kim, S. Oh, J. Park, and H. Kim, "A comparative study of time series anomaly detection models for industrial control systems," Sensors, vol. 23, no. 3, 1310, 2023.

[16]. A. Firouzi, S. Dadkhah, S. A. Maret, and A. A. Ghorbani, "DataSense: a real-time sensor-based benchmark dataset for attack analysis in IIoT with multi-objective feature selection," Electronics, vol. 14, 4095, 2025.

[17]. Y. Yang et al., "Kolmogorov–Arnold networks-based GRU and LSTM for loan default early prediction," Applied Soft Computing, vol. 197, 115213, 2026.

[18]. S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271, 2018.

[19]. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD, 2016, pp. 785–794.

[20]. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. 8th IEEE Int. Conf. Data Mining, 2008, pp. 413–422.

[21]. R. Rani et al., "Transformer, more than meets the eye: a deep learning approach to integrate rainfall time-series in multi-type landslide probability modeling," Geoscience Frontiers, vol. 17, issue 3, 102295, 2026.

[22]. Özalp, A. N., & Albayrak, Z. (2022). Detecting cyber attacks with high-frequency features using machine learning algorithms. *Acta Polytechnica Hungarica*, *19*(7), 213-233.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Boundary-Contact Modeling of Tumor Growth with Neural Surrogate Acceleration for Embedded Biomedical Systems

Manana Chumburidze
Akaki Tsereteli State University
Kutaisi, Georgia
manana.chumburidze@atsu.edu.ge
ORCID: 0000-0002-8029-9115

Teimuraz Sakhelashvili
Akaki Tsereteli State University
Kutaisi, Georgia
teimuraz.sakhelashvili@atsu.edu.ge
ORCID:0009-0006-1267-7045

Giorgi Chachua
Akaki Tsereteli State University
Kutaisi, Georgia
giorgi.chachua@atsu.edu.ge
ORCID: 0009-0003-6860-2403

Nona Kukhianidze
Akaki Tsereteli State University
Kutaisi, Georgia
nona.kukhianidze@atsu.edu.ge
ORCID: 0000-0001-9348-6289

Abstract
Tumor growth is strongly influenced by mechanical and anatomical constraints imposed by surrounding healthy tissue, which cannot be adequately described using standard boundary conditions. This paper presents a boundary-contact diffusion-reaction model for tumor growth that imposes unilateral contact constraints at the tumor-tissue interface. The model is formulated as a variational inequality and discretized using the finite element method to accurately capture constrained growth behavior. To reduce the high computational cost of repeated numerical simulations, a neural surrogate model is introduced to approximate the solution of the partial differential equation while preserving boundary and contact conditions. The surrogate is optimized for embedded-system deployment, enabling fast, energy-efficient tumor growth prediction on resource-constrained hardware. Numerical validation demonstrates that the proposed approach maintains accuracy in tumor boundary dynamics while significantly improving computational efficiency, making it suitable for real-time biomedical applications.

Keywords: Tumor, neural operators, embedded platforms,boundary-contact problem.

I. INTRODUCTION

Tumor growth occurs within a constrained biological environment, where interactions between expanding tumor tissue and surrounding healthy tissues play a critical role in disease evolution. Classical diffusion-reaction models with standard boundary conditions are often insufficient to capture these interactions, particularly when tumor expansion is mechanically or anatomically constrained. In this work, we present a boundary-contact mathematical framework for tumor growth modeling, formulated as a diffusion-reaction partial differential equation subject to unilateral contact constraints at the tumor-tissue interface. The proposed model distinguishes between freely evolving tumor boundaries and contact regions, imposing nonpenetration and flux-inequality conditions. These conditions naturally lead to a variational inequality formulation, enabling a mathematically consistent description of constrained biological growth. Spatial discretization is performed using the finite element method, while contact constraints are enforced using numerical strategies such as active-set and penalty-based methods [1]. Due to the high computational cost of repeatedly solving boundary-contact problems, especially in real-time or resource-constrained scenarios, we introduce a neural surrogate modeling approach. High-fidelity finite element simulations are used to train a constraint-aware neural network that approximates the solution operator of the underlying PDE while preserving essential boundary and contact properties. The surrogate model is optimized for embedded system deployment, addressing strict limitations on memory,

computation, and energy consumption typical of edge biomedical devices. Model compression and reduced-precision techniques are used to ensure fast, reliable inference on embedded hardware. Validation results demonstrate that the proposed surrogate-assisted framework achieves substantial reductions in computational time while maintaining accurate tumor boundary behavior and growth dynamics. The presented approach provides a promising foundation for real-time tumor growth prediction in embedded biomedical and clinical decision-support applications [2]. As a result, tumor evolution is not governed solely by bulk diffusion-reaction dynamics but is also strongly influenced by contact phenomena at the tumor-tissue interface. These interactions naturally lead to boundary-contact problems, where classical boundary conditions are insufficient to describe unilateral constraints such as:

- prevention of tumor penetration into rigid or semi-rigid surrounding tissue;
- growth inhibition at anatomical barriers;
- accumulation of tumor cells at the contact surface.

Boundary-contact formulations allow such effects to be modeled in a mathematically consistent and physically meaningful way, extending traditional boundary-value problems to include contact inequalities [3].

II. MATHEMATICAL MODEL

1. Governing Equation

Let us $\Omega(t) \subset R^n$ denote the tumor domain at time t, with boundary:

$$\partial\Omega(t) = \Gamma_free(t) \cup \Gamma_contact$$

where:

- Γ_free represents the freely evolving tumor boundary;
- $\Gamma_contact$ denotes the interface where the tumor is in contact with healthy tissue or anatomical constraints.

The diffusion-reaction equation governs the tumor cell density:

$$\frac{\partial u(x,t)}{\partial t} = \nabla \cdot \left(D \nabla u(x,t)\right) + f(u,c), \quad x \in \Omega(t) \tag{1}$$

where:

- $u(x,t)$ is the tumor cell density;
- D is the diffusion coefficient;
- $f(u,c)$ models proliferation, apoptosis, or necrosis, possibly dependent on oxygen or nutrient concentration c.

2. Boundary and Contact Conditions

On Γ_free, standard boundary or free-boundary conditions describe unconstrained tumor expansion. On the contact interface $\Gamma_contact$, unilateral constraints are imposed:

$$u \geq 0, \ \partial u/\partial n \leq 0, \ u(\partial u/\partial n) = 0 \quad on \ \Gamma_contact$$

These conditions express that:

- tumor density is non-negative;
- flux through the contact boundary cannot point outward into adjacent tissue;
- either tumor growth stops at the boundary, or the normal flux vanishes.

Such conditions are characteristic of boundary-contact unilateral problems and are well suited to modeling constrained biological growth processes [4]. From a biomedical perspective, the boundary-contact formulation captures the mechanical resistance of the surrounding tissue, saturation effects at tissue interfaces, clinically observed tumor deformation and restricted invasion, and growth patterns influenced by treatment zones, such as radiotherapy boundaries. This approach improves realism compared with purely free-boundary or Dirichlet/Neumann models by explicitly accounting for the mechanics of tumor-tissue interaction.

III. COMPUTATIONAL AND INTEGRATION

Boundary-contact conditions introduce nonlinearity and non-smooth behavior into the governing diffusion-reaction system, significantly increasing the computational cost of numerical simulations. The enforcement of unilateral contact constraints transforms the problem into a variational inequality, requiring iterative solution procedures and repeated updates of the active contact region. Consequently, high-fidelity finite element simulations may become computationally expensive, particularly for real-time prediction, treatment planning, and embedded biomedical applications. To mitigate this challenge, neural surrogate models, including Physics-Informed Neural Networks (PINNs) and related data-driven approaches, can be employed to approximate the solutions of the contact-constrained tumor growth problem. By incorporating the governing partial differential equation, boundary conditions, and contact constraints directly into the loss function, these models preserve the underlying physical principles while substantially reducing computational time. Such surrogate representations enable efficient prediction of tumor evolution without sacrificing consistency with the mathematical model [5]. To enforce unilateral contact constraints within the finite element framework, a primal-dual active-set strategy is adopted. At each iteration, boundary nodes are classified into active and inactive sets according to the current solution estimate. Nodes belonging to the active set satisfy the contact condition and are constrained, whereas inactive nodes evolve according to the governing diffusion-reaction equation.

IV. NUMERICAL IMPLEMENTATION

The governing tumor growth model, formulated as a diffusion-reaction partial differential equation with unilateral boundary-contact constraints, is discretized in space using the finite element method (FEM). A conforming mesh represents the computational domain, and piecewise-linear finite elements are employed to approximate the tumor density field. Time integration is performed incrementally, leading to a nonlinear algebraic system at each time step. To enforce the contact constraints, an **active-set strategy** is adopted. The method iteratively identifies regions where the contact condition is active and imposes the corresponding restrictions directly within the finite element formulation [6]. Starting from the current approximation u(k), the active and inactive sets are defined as

$$A=\{i:u_i(k)\leq tol\},\ I=\{i:u_i(k)>tol\}$$

where *tol* denotes a prescribed numerical tolerance.

The active set **AA** contains nodes for which the contact constraint is considered active. On these nodes, the tumor density is fixed according to the constraint condition

$$u_i=0,\ i \in A.$$

The finite element system is then solved only for the inactive nodes II, where the solution remains unconstrained. This produces an updated approximation u(k+1). Following each solution step, the active and inactive sets are recomputed and compared with those from the previous iteration. The iterative procedure continues until the active set remains unchanged between two successive iterations, indicating convergence of the contact region. The final solution satisfies both the governing diffusion-reaction equation and the imposed contact constraints.

1. Active-Set Iteration Algorithm for Contact Constraints

Input: Current approximation u(k), tolerance *tol*

1. Identify active nodes

A= {i:$u_i \leq$tol}

2. Identify inactive nodes

I={i:u_i>tol}

3. Impose contact constraints

$u_i=0,\ i \in A$

4. Solve the FEM system on the inactive set II
5. Compute the updated solution *u(k+1)*
6. Repeat until the active set no longer changes

Output: Contact-consistent solution you

This approach provides a robust and physically consistent treatment of tumor-tissue interaction while avoiding the need to introduce large penalty parameters. Moreover, the active-set strategy enables efficient identification of contact regions and can be incorporated into high-fidelity FEM simulations used for training neural surrogate models. In the hybrid framework, FEM solutions generated by the active-set procedure serve as reference data for machine learning models, including feedforward neural networks, physics-informed neural networks (PINNs), and neural operators [7]. Once trained, these models provide rapid approximations of tumor evolution while preserving the essential characteristics of the underlying PDE and contact mechanics. Figure 1. Schematic representation of tumor growth with and without contact constraints. In unconstrained regions, tumor cells proliferate freely, whereas growth is blocked at the contact boundary, resulting in localized stagnation and asymmetric tumor expansion near the contact interface.

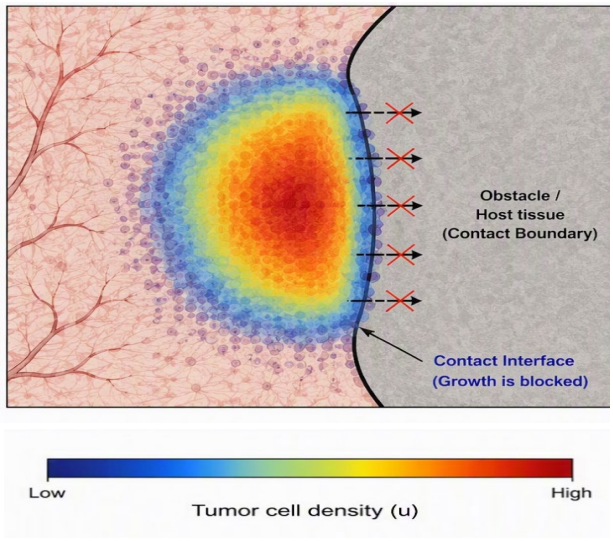


Figure 1. Effect of the Contact Constraint on Tumor Growth.

The numerical framework is validated by comparing surrogate predictions with reference FEM solutions for tumor density distribution, contact-region identification, and temporal evolution of tumor volume, thereby ensuring both numerical accuracy and computational efficiency.

V. EMBEDDED SYSTEM IMPLEMENTATION

For practical biomedical applications, including intraoperative decision support, portable diagnostic systems, and real-time monitoring platforms, tumor growth prediction models must operate under stringent computational and energy constraints. Although the finite element model with active-set contact enforcement provides accurate solutions, repeated numerical simulations remain computationally demanding and unsuitable for direct deployment on embedded hardware. To overcome this limitation, the proposed framework integrates a neural surrogate model trained on high-fidelity FEM simulations generated using the active-set algorithm. The target embedded platform is assumed to be a resource-constrained device, such as an ARM Cortex-M/A microcontroller, a low-power processor, or an edge AI system, characterized by limited computational capabilities, memory resources, and energy consumption [8]. Therefore, the neural surrogate is designed with a compact architecture to ensure low inference latency, predictable memory usage, and efficient execution under embedded-system constraints. To enable deployment on such platforms, model compression techniques including network pruning, weight quantization, and reduced-precision arithmetic may be applied while maintaining acceptable prediction accuracy. The embedded model approximates the mapping from tumor growth parameters to the corresponding contact-consistent solution obtained via the FEM active-set procedure.

The input features typically include:

- geometric characteristics of the tumor domain and boundary;
- diffusion and reaction parameters of the tumor growth model;
- descriptors related to contact regions and boundary constraints.

The outputs may consist of:

- tumor cell density at selected spatial locations;

- reconstructed tumor concentration fields;
- tumor volume and invasion-front indicators;
- information associated with active contact regions.

Contact effects are represented implicitly through training datasets generated from FEM simulations employing the active-set algorithm. As a result, the neural model learns the characteristic behavior associated with constrained tumor growth near contact boundaries and reproduces the corresponding solution patterns without explicitly solving the optimization problem during inference.

At runtime, the embedded system executes the following lightweight computational workflow:

1. acquisition of patient-specific or simulation parameters;
2. preprocessing and normalization of input data;
3. neural surrogate inference on embedded hardware;
4. post-processing to ensure physical admissibility, including non-negativity of tumor density;
5. generation of tumor growth indicators and predictive outputs.

Because only a forward neural-network evaluation is required, the computational burden is substantially reduced compared with repeatedly executing the active-set FEM solver. This enables near real-time prediction of tumor evolution while preserving the essential characteristics of the underlying diffusion-reaction-contact model. To evaluate reliability, embedded predictions are compared with reference FEM solutions obtained using the active-set algorithm. Validation metrics include spatial tumor-density distributions, identification of contact regions, preservation of constrained growth behavior near boundaries, and temporal evolution of tumor volume. The comparison demonstrates that the surrogate-assisted embedded implementation maintains consistency with the underlying mathematical model while significantly reducing computational complexity and execution time. Consequently, the proposed embedded framework provides an efficient and practical solution for biomedical applications requiring rapid tumor-growth prediction under contact constraints, while retaining the physical interpretation and numerical characteristics of the original FEM active-set formulation.

VI. CONCLUSION

This paper presented a boundary-contact mathematical framework for modeling tumor growth in mechanically and anatomically constrained biological environments. By formulating tumor evolution as a diffusion-reaction partial differential equation with unilateral contact conditions, the proposed model captures important tumor-tissue interaction mechanisms that are not represented in conventional boundary formulations. The numerical solution was obtained using a finite-element discretization combined with an active-set strategy to enforce contact constraints. The active-set procedure iteratively identifies contact and non-contact regions and ensures physically consistent satisfaction of the imposed boundary conditions. This approach provides an effective and robust mechanism for modeling constrained tumor growth while preserving the mathematical structure of the governing problem. To address the high computational cost of repeated FEM simulations, a neural surrogate modeling framework was introduced and adapted for deployment on embedded systems. The surrogate model was trained using high-fidelity FEM solutions generated with the active-set algorithm and was designed to reproduce tumor growth behavior and contact-region dynamics with significantly reduced computational requirements. The results indicate that the surrogate-assisted framework substantially improves computational efficiency while maintaining good agreement with reference FEM solutions for tumor density distribution, constrained boundary behavior, and overall tumor growth dynamics. Consequently, the proposed approach offers a practical foundation for real-time tumor growth prediction on resource-constrained embedded platforms.

REFERENCES

[1]. C. Wessels, C. Böhm, F. Aldakheel, M. Hüpgen, M. Haist, L. Lohaus, and P. Wriggers, "Computational homogenization using convolutional neural networks," in Current Trends and Open Problems in Computational Mechanics, Springer, 2022, pp. 569-579.

[2]. M. Uteuliyeva, A. Zhumekenov, R. Takhanov, Z. Assylbekov, A. J. Castro, and O. Kabdolov, "Fourier neural networks: A comparative study," Intelligent Data Analysis, vol. 24, no. 5, pp. 1107-1120, 2020.

[3]. A. Climescu-Haulica, "Broad research in artificial intelligence and neuroscience," BRAIN. Broad Research in Artificial Intelligence and Neuroscience, vol. 15, no. 4, pp. 375-386, 2024.

[4]. M. Chumburidze and V. Niminet, "Efficient modelization for complex geometries of tumor growth via thermo-elastic diffusion partial differential equations and artificial neural networks," BRAIN. Broad Research in Artificial Intelligence and Neuroscience, vol. 16, no. 1, pp. 315-323, 2025.

[5]. M. Chumburidze, M. Mnatsakaniani, D. Lekveishvili, and N. Julakidze, "Numerical study of dynamical system using deep learning approach," Open Journal of Applied Sciences, vol. 15, no. 2, pp. 425-432, 2025.

[6]. D. Nayak and S. Goswami, "Data-efficient time-dependent PDE surrogates: Graph neural simulators vs neural operators," arXiv preprint arXiv:2509.06154, 2025.

[7]. M. Laurie and J. Lu, "Explainable deep learning for tumor dynamic modeling using Neural-ODE," 2023.

[8]. TumorNetSolvers, "Deep learning framework for Fisher-Kolmogorov PDE-based tumor growth modeling," 2024.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Decision Tree Based Data Completion in Structured Lexicographic Resources

Archil Elizbarashvili
Computer Science Department of the Faculty of Exact and Natural Sciences of Ivane Javakhishvili Tbilisi State University
Tbilisi, Georgia
archil.elizbarashvili@tsu.ge
ORCID:0000-0002-1673-0951

Magda Tsintsadze
Computer Science Department of the Faculty of Exact and Natural Sciences of Ivane Javakhishvili Tbilisi State University
Tbilisi, Georgia
magda.tsintsadze@tsu.ge
ORCID:0000-0001-9316-3295

Abstract
This paper presents a Decision Tree-based approach for completing missing linguistic information in KARTUVERBS, a large-scale linked-data resource for Georgian verb lemmatization and translation. The study focuses on reconstructing missing verbal nouns, which are essential for linking inflected verb forms to their corresponding lemmas. The task is formulated as a supervised multi-class classification problem, in which the model predicts missing verbal nouns based on structured morphological and paradigm-level attributes. Five-fold KFold evaluation achieved a mean accuracy of 99.93%, demonstrating highly stable performance when related lexical and paradigm groups were represented in the training data. A stricter GroupKFold evaluation produced a mean accuracy of 76.91%, showing that prediction is more difficult for entirely unseen paradigms. Per-class results vary for rare and ambiguous classes, mainly due to class imbalance and the lexeme-specific nature of Georgian verbal morphology. These findings demonstrate that the Decision Tree model is particularly well-suited to completing missing entries in an existing lexicographic resource, especially when related lexical and paradigm patterns are already represented in the training data. Overall, the results indicate that Decision Trees provide an effective, interpretable, and computationally efficient method for completing structured lexicographic data in morphologically rich and low-resource language contexts.

Keywords: Decision Tree, Machine Learning, Data Completion, Lexicographic Resources, NLP

I. INTRODUCTION

Supervised machine learning uses labeled data, where each training example is associated with a known target value. By learning relationships between input features and target classes, a model can apply these patterns to new, previously unseen cases [1],[2]. A typical workflow includes data cleaning, structuring, feature selection, model training, and evaluation [1]. After preprocessing, the dataset is divided into training and test subsets: the training subset is used to build the model, while the test subset is used to evaluate generalization [2]. This distinction between learning and generalization is essential. A model that performs well only on the training data may reproduce the training examples instead of learning reusable patterns. Therefore, evaluation on independent data is necessary for assessing whether a machine-learning approach is suitable for practical use [3]. A central issue is the balance between model simplicity and complexity, often described by the bias-variance trade-off [4].

An overly simple model may underfit the data, while an overly complex model may overfit by adapting too closely to noise or accidental regularities. The desired model should capture meaningful patterns while avoiding excessive dependence on the training set. Decision Tree algorithms are particularly relevant in this context because they provide an interpretable way of modeling classification decisions [5]. Unlike many black-box approaches, a Decision Tree represents the learned process as a sequence of feature-based rules. This makes the model useful not only for prediction, but also for explaining why a particular class was assigned. Such interpretability is especially valuable in linguistic and lexicographic tasks, where model decisions should remain understandable and verifiable. The present study investigates the use of a Decision Tree classifier for completing missing linguistic data in a large-scale Georgian verb resource. Georgian verb forms contain rich morphological information [8], [9], [10], and some lemma-related attributes may be absent or incomplete in the original data. The task, therefore, requires a model that can learn from existing annotated examples and infer missing values from combinations of morphological features. The following sections present the KARTUVERBS use case [6], describe the Decision Tree algorithm, and discuss the experimental results obtained for reconstructing missing verbal nouns.

II. BACKGROUND AND MOTIVATION

1. A Linguistic resource

KARTUVERBS is a linked-data-based system developed for lemmatizing and translating Georgian verbs [7]. Its main objective is to connect Georgian inflected verb forms with their lemmas [13], [14], morphological attributes, paradigmatic information, and external lexical resources [15], [16]. This is important because Georgian has complex verbal morphology and no standardized infinitive comparable to English's. KARTUVERBS relies on data from the CLARINO/INESS infrastructure, made available thanks to Paul Meurer [8]. These data ultimately derive from Kita Tschenkeli's Georgian-German Dictionary [9]. The original resource is organized as a tree-like hierarchy indexed mainly by a common verb root. Although useful for lexicographic representation, this structure is less convenient when the goal is to retrieve information starting from a concrete inflected form. In the original dataset, each Georgian verb is represented through many inflected forms distributed across 11 tense groups and 6 grammatical persons. More than 60 features or attributes describe each verb. The data, therefore, contain rich linguistic information, but also heterogeneous attributes and incomplete fields. One important problem identified during the construction of KARTUVERBS was the absence of lemma-related information. In particular, some inflected verb forms did not have the corresponding verbal noun filled in consistently. This created an obstacle for automatic form-to-lemma navigation and for the construction of a complete semantic graph. During construction, the data were extracted, cleaned, filtered, and transformed into a relational representation before conversion into RDF [19], [20], [21]. This changed the resource's access logic: instead of starting from a root, users and systems can start from any inflected form or morphological attribute and move toward the associated lemma, paradigm, or grammatical category. The resulting resource contains approximately five million inflected forms corresponding to more than 16,000 Georgian verbs, with more than 80 million semantic relations. At this scale, manually completing missing fields is not realistic. Machine learning, therefore, becomes a practical tool for improving data completeness [10] before generating the final semantic graph [11]. The task is not open-ended text generation, but supervised classification over a controlled set of possible linguistic values.

2. Motivation for choosing Decision Trees

A Decision Tree model was selected for the KARTUVERBS data-completion task because the dataset is highly structured and mostly categorical. It contains grammatical and morphological attributes, including tense, person, number, preverb, root, suffix, morphological type, and paradigm-related identifiers. Since the output classes belong to a defined linguistic inventory, classification is more appropriate than generative modeling.

Interpretability was another important motivation. In lexicographic and linguistic resources, it is not sufficient for a model to produce a prediction; researchers should also understand why it was made. A Decision Tree provides a transparent structure where each path from the root to a leaf can be read as a rule and compared with grammatical expectations. The use of a Decision Tree is also justified by the structure of Georgian verb morphology. Georgian verbs may include preverbs, stem changes, version markers, voice-related features, and other morphological components. These features interact in systematic ways. A Decision Tree can capture such interactions without requiring a manually written rule for every possible case.

Compared with more complex or less transparent models, the Decision Tree approach offers a useful balance between performance, simplicity, and explainability. In preliminary experiments, SVM performed worse, and Random Forest was not retained because it required substantially more computational resources. In the KARTUVERBS case, the goal is not to model unrestricted language use, but to recover missing fields in a controlled lexicographic dataset. The following section presents the algorithm in more detail.

III. DECISION TREE ALGORITHM

A Decision Tree is a supervised machine-learning algorithm mainly used for classification. It assigns each input instance to one of several predefined classes by learning from labeled examples. The learned decision process is represented as a tree [12]. A decision tree consists of a root node, internal decision nodes, branches, and leaf nodes. The root node contains the initial dataset; internal nodes correspond to tests on selected features; branches represent possible outcomes; and leaf nodes provide the final class prediction. Each path from the root to a leaf can therefore be interpreted as an if-then rule. The tree is built through recursive partitioning. At each node, the algorithm selects the feature that best separates the data into more homogeneous subsets. This process continues until a stopping condition is reached. Several decision-tree algorithms exist, including ID3, C4.5, C5.0, and CART. In this work, the scikit-learn implementation is based on CART, which constructs binary trees and commonly uses the Gini Index as a classification criterion [13].

The quality of a split is evaluated using impurity measures. One common measure is the Gini Index:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{1}$$

where D is the dataset, m is the nnumber of classes, and p alt i. is the probability that a randomly selected instance belongs to class i, with p alt i. eşittir mutlak değer büyük harf D alt i.bclasses, and p_i is the probability that a randomly sfrom -1 to 1ce belongs to class i, with $p_i = \frac{|D_i|}{|D|}$. The Gini Index ranges in $\left[0, 1 - \frac{1}{m}\right]$. If $Gini(D) = 0$, the dataset is perfectly homogeneous, meaning that all instances belong to one class. The maximum value is attained when classes are evenly distributed.

Another widely used measure of impurity is entropy. Entropy measures the uncertainty or information content of a dataset and is defined as:

$$Entropy(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{2}$$

If $Entropy(D) = 0$, the dataset is fully homogeneous. The maximum entropy, $log_2(m)$, is attained when classes are evenly distributed.

A good split is one that reduces entropy, because it produces subsets that are more ordered and more class-specific than the original dataset. The reduction of entropy after splitting a dataset by a feature is called Information Gain. For a dataset D and an attribute A, Information Gain is computed as:

$$IG(D,A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Entropy(D_v) \quad (3)$$

where $Entropy(D)$ is the initial entropy, $Values(A)$ is the set of possible values of attribute A, and D_v is the subset for which $A = v$. The factor $\frac{|D_v|}{|D|}$ represents the weight of subset D_v within the full dataset. At each node, the tree selects the attribute with the highest Information Gain. Analogously, Gini Gain measures how much the Gini Index decreases after the split:

$$GG(D,A) = Gini(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Gini(D_v) \quad (4)$$

Both criteria follow the same principle: a good split transforms a mixed dataset into smaller and more homogeneous subsets.

The recursive splitting process stops when the subset is sufficiently homogeneous, no useful features remain, the subset is too small, the maximum depth is reached, or further splitting no longer improves classification. These stopping conditions are important because an unrestricted tree may overfit the training data. Parameters such as maximum depth, minimum samples per split, and minimum samples per leaf can reduce this risk.

Decision Trees are useful for linguistic data completion because they are interpretable, can handle numerical and categorical features, require limited preprocessing, and naturally perform feature selection. Their main limitations are sensitivity to small changes in data and reduced interpretability when the tree becomes too deep. To illustrate the interpretability of the Decision Tree approach, Figure 1 presents a simplified Decision Tree with a maximum depth of three. The model first divides the data according to the encoded verb root and subsequently uses the verb identifier, Tschenkeli class, and sub-paradigm identifier. These variables are among the most important features in the complete model, indicating that verbal-noun prediction depends mainly on lexical identity and paradigm-level structure.

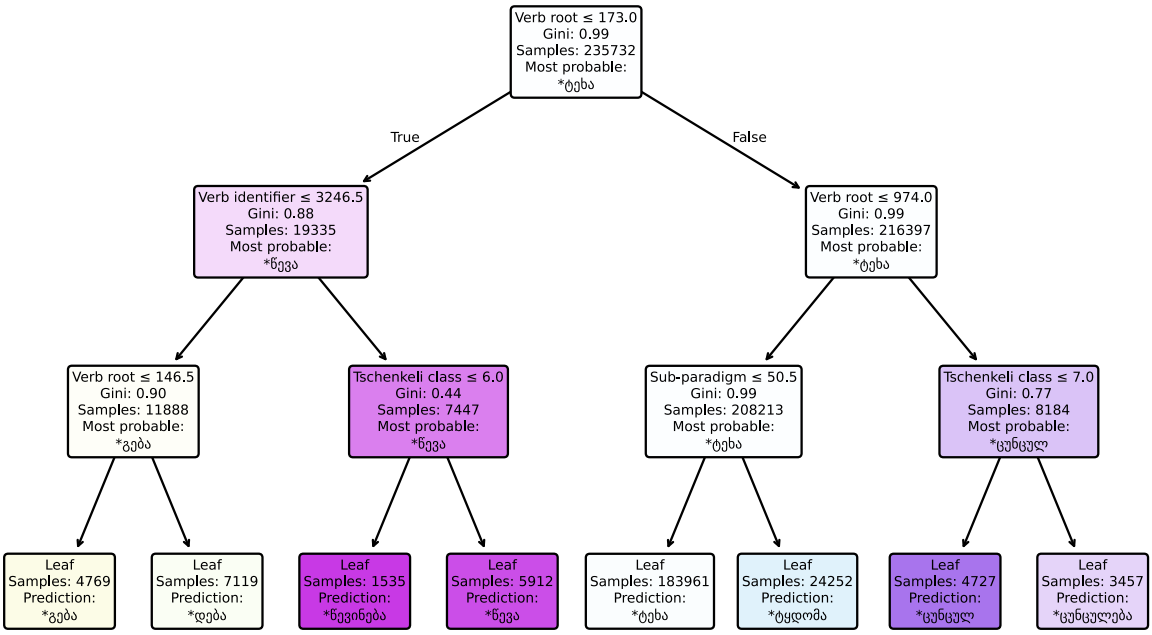


Figure 1. Simplified Decision Tree with encoded splitting rules and predicted verbal nouns.

Figure 1 contains eight terminal leaves representing eight possible prediction outcomes. Each input record follows only one branch at every decision node and reaches one final leaf. Internal nodes display the splitting condition, Gini impurity, number of training instances, and the most frequent verbal noun among the records reaching that node. The "Most probable" value shown in internal nodes indicates the most frequent verbal noun at that stage, not the final output. The final verbal noun is shown only in the terminal leaf reached by the input record. Because the attributes were numerically encoded, the thresholds shown in the figure represent internal computational values rather than direct linguistic measurements. The simplified tree is provided for visualization and interpretability, whereas the reported performance results were obtained using the complete Decision Tree model. In the KARTUVERBS use case, the Decision Tree is appropriate because the task is a supervised multi-class classification problem. The input data consist of morphological and grammatical attributes of Georgian verb forms, and the target value is a missing verbal noun. Since the possible outputs belong to a controlled linguistic inventory, classification is more suitable than open-ended generation.

IV. EXPERIMENTAL SETUP

The experiments were conducted in Python using Anaconda Navigator and Jupyter Notebook. The Decision Tree model was implemented with the *DecisionTreeClassifier* class from the *scikit-learn* library. This classifier is based on CART and constructs a binary tree. By default, it uses the Gini impurity criterion, although entropy-based splitting can also be selected. Before training, the dataset was filtered to focus on the reconstruction task. Only cases were retained in which the verbal noun was not consistently present across all paradigms within a common verb-root group. This reduced the data to approximately 300,000 rows. These rows represented cases where the target variable, the verbal noun, was partially missing and had to be inferred from other linguistic features. The task was formulated as a supervised multi-class classification problem. The model was trained on records in which the verbal noun was known, and then used to predict missing verbal-noun values in incomplete records. Since the outputs belong to a controlled inventory of known verbal nouns, the task is classification

rather than open-ended generation. The model used 14 core features selected from morphological and structural attributes, including tense, person, number, verb root, preverb, marker-12, voice/theme marker, ending, common verb-root identifier, morphology type, class, and contact marker. The target verbal noun was excluded from the input feature matrix before model training, thereby preventing direct target leakage. Most input attributes were categorical and were converted into numerical values using label encoding. Georgian text fields required additional processing because Georgian characters are represented in UTF-8 by multiple bytes [26], [27], [28]. Therefore, each Georgian word was encoded as a single numerical value by extracting the last byte of each Georgian character's UTF-8 representation and summing these byte values. This produced a compact representation and reduced computational cost. Because this encoding is sum-based, different strings may theoretically receive the same numerical value. Anagrams necessarily produce the same sum, while other distinct strings may also collide. However, predictions were not based on a single encoded field: the model simultaneously used lexical, morphological, and paradigm-level attributes, including tense, person, root, preverb, formants, morphology type, and structural identifiers. The target variable was encoded separately: each distinct verbal noun received a unique numerical identifier, with a mapping table that preserved the correspondence between the encoded labels and the original linguistic values. This made it possible to decode predictions back into meaningful verbal nouns. A clear distinction was also made between zero values and missing values. Zero values were preserved as meaningful categories, while missing values were treated as values to be predicted. The dataset was divided into training and test subsets using an 80%-20% split. To reduce dependence on a single random split, more than ten different random seeds were tested. The results remained practically unchanged, indicating that the model was stable. In addition to the hold-out evaluation, KFold cross-validation was used to assess stability across different subsets. A stricter GroupKFold evaluation was also performed to ensure that all forms belonging to the same verbal paradigm were kept entirely within either the training or test fold. The model was evaluated using accuracy, precision, recall, and F1-score. Because the task involved multiple classes with an uneven distribution, both macro-averaged and weighted-averaged metrics were considered.

V. RESULTS AND DISCUSSION

The objective of the experiment was to reconstruct missing verbal nouns in KARTUVERBS using a Decision Tree classifier. The model achieved approximately 99% accuracy on the test set, which is a strong result for a multi-class classification problem involving many verbal-noun classes.

Table 1. Overall classification performance using 14 input features.

	Precision	**Recall**	**F1-score**	**Support**
Accuracy			0.99	58933
Macro avg	0.98	0.98	0.98	58933
Weighted avg	0.99	0.99	0.99	58933

The overall classification report is presented in Table 1. The test set contains 58,933 evaluated instances. Accuracy is 0.99, while macro-averaged and weighted-averaged Precision, Recall, and F1-score are in the range of 0.98-0.99. In single-label multi-class classification, micro-averaged Precision, Recall, and F1-score are equivalent to Accuracy, because each instance has one true class and one predicted class. Although the aggregate results are high, they do not fully reflect performance across all verbal-noun classes. Weighted metrics are influenced primarily by frequent classes, whereas rare classes may have unstable or substantially lower scores. A per-class analysis is therefore necessary to identify difficult cases.

Table 2. Selected individual per-class verbal-noun prediction results.

	Precision	Recall	F1-score	Support
…	…	…	…	…
24	0.83	0.86	0.84	140
277	0.00	0.00	0.00	10
278	0.00	0.00	0.00	22
279	0.71	0.92	0.80	84
280	0.59	1.00	0.74	89
282	0.00	0.00	0.00	23
283	0.00	0.00	0.00	39
284	1.00	0.84	0.91	255
285	0.00	0.00	0.00	58
288	1.00	0.43	0.6	30
289	0.00	0.00	0.00	13
290	0.00	0.00	0.00	23
291	0.00	0.00	0.00	26
292	1.00	0.74	0.85	111
418	1.00	0.99	1.00	112
420	0.99	1.00	1.00	142
449	0.58	0.53	0.55	53
462	0.96	1.00	0.98	26
467	0.67	0.80	0.73	51
468	0.40	0.98	0.56	140

Table 2 shows that performance varies across individual verbal-noun classes. Some classes are predicted almost perfectly, while others have lower scores. In particular, for classes such as 277 corresponding to *ღწევა[1] (“to reach”, „to achieve”, „to escape”), Precision, Recall, and F1-score are all 0.00. The support value of 10 indicates that the test set contains 10 actual instances of class 277, while a Recall of 0.00 means none were correctly identified. A Precision of 0.00 may indicate either that the model never predicted this class, in which case Precision is mathematically undefined and was reported as 0.00 using zero_division=0, or that the model predicted the class but all such predictions were incorrect. In both cases, the absence of true-positive predictions results in an F1-score of 0.00. These zero scores mainly occur in low-support or difficult classes and have limited influence on the weighted average, as they represent only a small proportion of the test set.

Other classes illustrate different error patterns. For class 288, corresponding to *ყრდნობა („to lean”, “to rely”, “to support”), Precision is 1.00 and Recall is 0.43, with a support of 30. This means that every prediction assigned to class 288 was correct, but the model identified only approximately 43% of its actual instances, or about 13 out of 30. In contrast, class 468, corresponding to *ყოლა (“to have” or “to follow”) has a Precision of 0.40 and

[1] An asterisk at the beginning of a verb indicates that it may be replaced by different preverbs. The resulting prefixed forms may have different meanings.

a Recall of 0.98, with a support of 140. The model, therefore, identified almost all actual instances of class 468, but also incorrectly assigned many instances from other classes to this label.

This variation is expected in a large multi-class linguistic classification task, especially when the class distribution is uneven. Rare classes have less stable Precision, Recall, and F1-score values, because even a few incorrect predictions can significantly affect their scores. The linguistic nature of Georgian verbs also contributes to this variation. Georgian verbal morphology includes tense, person, number, preverbs, thematic markers, contact markers, stem alternations, and paradigm-level regularities. Some verbal nouns follow predictable patterns, while others are more lexeme-specific or ambiguous. The results support the choice of a Decision Tree classifier. KARTUVERBS contains structured, numerically encoded morphological, lexical, and grammatical attributes. Decision Trees can divide this feature space into interpretable decision paths, which is particularly valuable in lexicographic work, where predictions should remain inspectable and linguistically verifiable. Since the task includes many verbal-noun classes, most of which are predicted with very high accuracy, the complete confusion matrix appears almost entirely diagonal and does not clearly reveal the remaining errors.

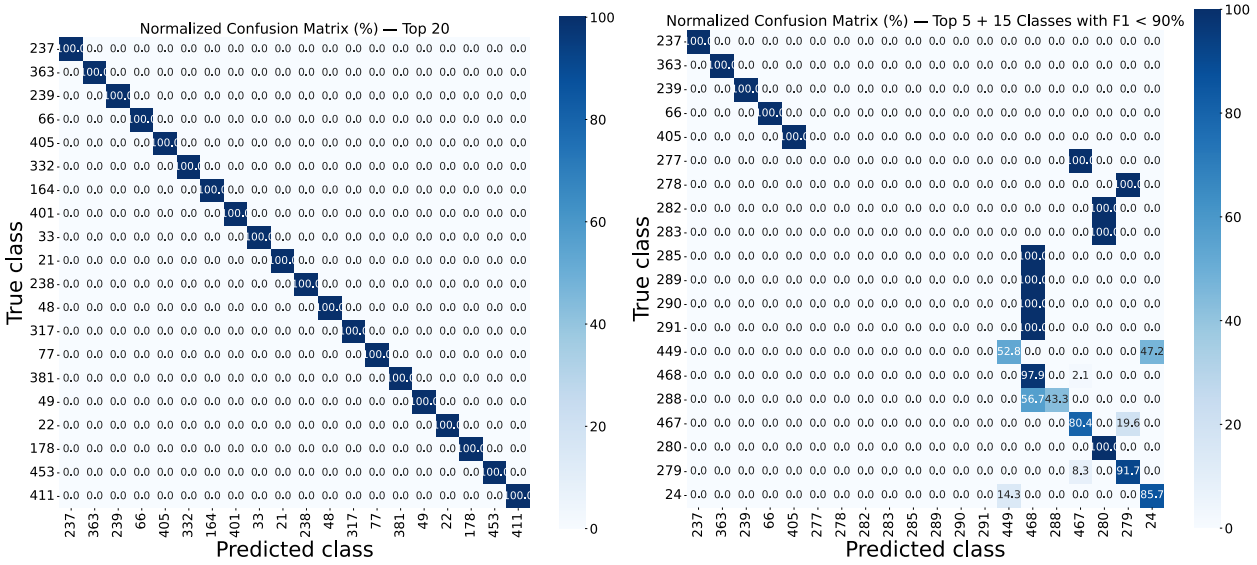


Figure. 2. Normalized confusion matrices for frequent and lower-performing verbal-noun classes.

Figure 2 therefore presents normalized confusion matrices for selected classes. The left matrix shows the 20 most frequent classes, most of which achieve near-100 % accuracy. The right matrix combines the five most frequent classes with 15 additional classes whose F1-scores are below 90%. This representation makes it possible to compare highly stable classes with more difficult cases and shows that most classification errors are concentrated in a relatively small number of less frequent or ambiguous classes.

The reliability of the results was assessed using both KFold and GroupKFold cross-validation. In five-fold KFold evaluation, the test accuracy ranged from 99.92% to 99.95%, with a mean of 99.93% and a standard deviation of 0.015 percentage points. The mean training accuracy was approximately 99.97%. These highly consistent results indicate that the model is stable when the training and test folds contain forms drawn from the same overall inventory of lexical and paradigm groups. A stricter evaluation was performed using GroupKFold, in which all records belonging to the same verbal paradigm were assigned entirely to either the training or test fold. No group overlap occurred in any of the five folds. Under this setting, test accuracy ranged from 74.74% to

78.68%, with a mean of 76.91% and a standard deviation of 1.46 percentage points. Training accuracy remained approximately 99.96%.

The difference between the KFold and GroupKFold results reflects the strongly lexeme- and paradigm-dependent nature of Georgian verbal-noun prediction. Standard KFold allows forms from the same paradigm to appear in both training and test folds and therefore measures the model's ability to reconstruct missing values within lexical groups already represented in KARTUVERBS. GroupKFold instead evaluates generalization to entirely unseen paradigms. The lower GroupKFold score consequently indicates limited out-of-group generalization rather than direct target leakage. The target verbal noun was excluded from the input features, and the absence of group overlap confirms that complete paradigms were strictly separated during the GroupKFold evaluation.

These findings clarify the intended scope of the method. The Decision Tree is highly effective for completing fragmentarily missing verbal nouns when related lexical or paradigm information is already present in the resource. Still, it is less reliable for predicting verbal nouns for completely unseen paradigms.

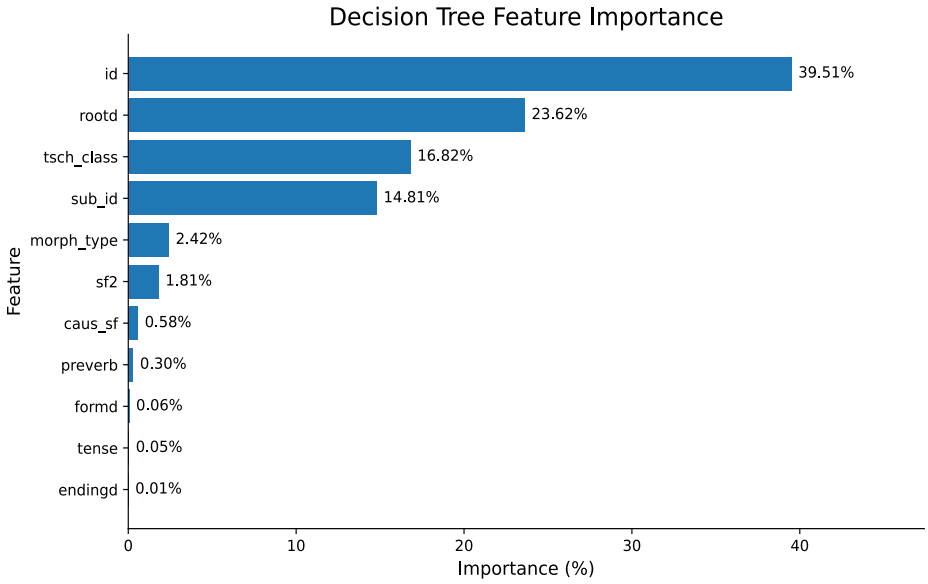


Figure 3. Decision Tree features importance

The feature-importance analysis also provides useful insight into the model's behavior. The most influential feature was the verb identifier, which accounted for approximately 39.51% of the total importance. It was followed by the verb root, with 23.62%, the Tschenkeli class, with 16.82%, and the sub-paradigm identifier, with 14.81% (see Figure 3). Other morphological attributes, such as morphology type, stem formant, causative stem formant, and preverb, had lower individual importance scores. These results suggest that the model relies primarily on lexical and paradigm-level information, while still using morphological features to refine the prediction.

At the same time, this result should be interpreted in light of the approach's main limitations. The source data contained more than 60 characteristics, while the model was trained on a reduced set of 14 selected features. This assumes that these selected attributes contain sufficient information for reliable prediction; however, this assumption was checked experimentally through model evaluation and validation. A second limitation is that the method can restore missing values only when the necessary regularities are already present in the existing

data, making it more suitable for paradigm completion than for predicting completely unknown lexical patterns. From a practical point of view, the approach is valuable because manually completing missing verbal nouns in a resource of this size is not feasible. A machine-learning model can automatically reconstruct most missing values, while rare or lower-confidence cases can be reserved for expert validation. Compared with more complex models, the Decision Tree offers a useful balance between accuracy, simplicity, computational efficiency, and explainability. The experiment shows that the Decision Tree classifier is effective for reconstructing missing verbal nouns in KARTUVERBS. Although per-class performance varies, this is expected due to class imbalance and the linguistic complexity of Georgian verbal morphology. The results confirm that the method is reliable for completing structured lexicographic resources when the objective is to recover missing information within known paradigms. As future work, we plan to use the crowdsourcing platform for human-in-the-loop validation of questionable or low-confidence verbal noun predictions [14]. This validation will focus on rare and ambiguous cases, helping to improve the reliability and quality of the completed KARTUVERBS data.

VI. CONCLUSION

This paper presented a Decision Tree-based approach for completing missing verbal nouns in KARTUVERBS, a linked-data resource for Georgian verb lemmatization and translation. The task was formulated as a supervised multi-class classification problem, in which the model predicts missing values based on structured morphological and paradigm-level attributes. The model achieved a mean KFold accuracy of 99.93%, demonstrating reliable completion within lexical and paradigm groups already represented in the resource. The lower GroupKFold accuracy of 76.91% indicates that prediction is more difficult for paradigms completely unseen, reflecting the lexeme-dependent nature of Georgian verbal morphology. Overall, the Decision Tree approach provides a practical balance between accuracy, efficiency, and interpretability. It is suitable for enriching incomplete lexicographic data before conversion into a linked-data semantic graph, while rare or ambiguous cases can be further validated through expert review and crowdsourcing.

ACKNOWLEDGEMENT

This research was supported by the Shota Rustaveli National Science Foundation of Georgia and the Franco-Georgian University. The paper is based on the author's doctoral research conducted at Ivane Javakhishvili Tbilisi State University.

REFERENCES

[1]. S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," Machine Learning, vol. 16, pp. 235-240, 01 September 1994.
[2]. J. Watt, R. Borhani, A. K. Katsaggelos and C. U. Press, Machine Learning Refined: Foundations, Algorithms, and Applications, Cambridge University Press, 2020.
[3]. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, March 2021.
[4]. B. Thomas and A. K. John, "Machine Learning Techniques for Recommender Systems – A Comparative Case Analysis," IOP Conference Series: Materials Science and Engineering, vol. 1085, p. 012011, February 2021.
[5]. C. Molnar, Interpretable Machine Learning, 3rd ed., 2025.
[6]. E. Briscoe and J. Feldman, "Conceptual Complexity and the Bias/Variance Tradeoff," Cognition, vol. 118, pp. 2-16, January 2011.

[7]. M. Bansal, A. Goyal and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," Decision Analytics Journal, vol. 3, p. 100071, 2022.

[8]. M. Cherchi, Modern Georgian Morphosyntax: A grammatico-categorial hierarchy-based analysis with special reference to indirect verbs and passives of state, Harrassowitz Verlag · Wiesbaden, 1997.

[9]. D. Melikishvili, J. D. Humphries and M. Kupunia, The Georgian Verb: A Morphosyntactic Analaysis, Dunwoody Press, 2010.

[10]. K. Tuite, Kartvelian morphosyntax: Number agreement and morphosyntactic orientation in the South Caucasian languages, Lincom Europa Munich, 1998.

[11]. M. Ducassé and A. Elizbarashvili, "Finding Lemmas in Agglutinative and Inflectional Language Dictionaries with Logical Information Systems: The Case of Georgian verbs," in Proceedings of XX EURALEX International Congress, 2022.

[12]. Elizbarashvili, M. Ducassé, M. Khachidze and M. Tsintsadze, "From Structured Textual Data to Semantic Linked-data for Georgian Verbal Knowledge," in Proceedings of the eLex 2023 conference, 2023.

[13]. J. Gippert, "Complex Morphology and its Impact on Lexicology: the Kartvelian Case," in Proceedings of the 17th EURALEX International Congress, Tbilisi, 2016.

[14]. T. Margalitadze, "The Lexicography of Georgian," in International Handbook of Modern Lexis and Lexicography, P. Hanks and G. de Schryver, Eds., Springer Berlin Heidelberg, 2020, p. 1–24.

[15]. E. M. Bender, "Linguistically naïve != language independent: why NLP needs linguistic typology," in Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?, USA, 2009.

[16]. J. DeCesaris, "Dictionaries and Morphology," in Proceedings of XIX EURALEX International Congress, 2021.

[17]. P. Meurer, "A computational grammar for Georgian," in International Tbilisi Symposium on Logic, Language, and Computation, 2007.

[18]. K. Tschenkéli, Y. Marchev and L. Flury, Georgisch-deutsches Wörterbuch, vol. 2, Amirani-Verlag Zürich, 1965.

[19]. J. Gracia, I. Kernerman and J. Bosque-Gil, "Toward Linked Data-native Dictionaries," in Proceedings of the eLex 2017: Lexicography from Scratch, Netherlands, 2017.

[20]. W3C, "RDF 1.1 Concepts and Abstract Syntax," 2014.

[21]. T. Berners-Lee, Semantic Web Road map, 1998.

[22]. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., New, York: Springer Science & Business Media, 2009.

[23]. D. Lindemann, C. Klaes and P. Zumstein, "Metalexicography as Knowledge Graph," in Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019), 2019.

[24]. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, "Classification and Regression Trees," 1984.

[25]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[26]. Elizbarashvili, M. Tsintsadze and M. Khachidze, "Enhancing Georgian Text Processing: Transliteration Techniques," Baltic Journal of Modern Computing, vol. 12, p. 415–430, 2024.

[27]. ISO/IEC, "ISO/IEC 10646:2020, Information Technology — Universal Coded Character Set (UCS)," 2020.

[28]. Unicode Consortium, "Components of the Unicode® Standard," 2024. [Online]. Available: https://www.unicode.org/versions/components-15.1.0.html.

[29]. D. Gross-Amblard, M. Tommasi, I. Rakotoniaina, C. Thierry, R. Singh and L. Jacoboni, "HEADWORK: Powering the Crowd with Tuple Artifacts," INRIA.HAL.SCIENCE Web Portal, 2023.



International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Comparative Analysis of Vision Transformer Models for Multimodal Emotion Recognition

İlya KUŞ
Gazi University, Graduate School of Natural and Applied Sciences
Ankara, Türkiye
ilya.kus@gazi.edu.tr
ORCID: 0000-0001-5850-949X

Cemal KOÇAK
Gazi University, Faculty of Technology, Department of Computer Engineering
Ankara, Türkiye
ccckocak@gazi.edu.tr
ORCID: 0000-0002-8902-0934

Ayşe KELES
School of Computer Science, University of Galway, Galway, Ireland
University
Ankara, Türkiye
ayse.keles@universityofgalway.ie
ORCID: 0000-0001-8760-412X

Abstract
Emotion recognition is a crucial research area for human-computer interaction, signal processing, and intelligent decision support systems. Human emotions are not only reflected in textual content but also expressed through tone of voice, emphasis, facial expressions, and visual behaviors. Therefore, multimodal approaches offer a more accurate representation than unimodal methods. In this study, the performance of different Vision Transformer (ViT)- based architectures was comparatively examined in the context of multimodal emotion recognition. In the experimental setup, the text modality was represented using DistilBERT, the audio modality was represented with an MFCC-based BiGRU-attention structure, and architectures such as BEiTv2, CaFormer, CoAt, ConViT, DeiT3, FastViT, MobileViT, MViTv2, NextViT, PiT, PVTv2, Swin Transformer, and TinyViT were used for the visual modality. By keeping the text and audio modalities constant, the impact of changing only the visual modality architecture on multimodal emotion recognition performance was analyzed. Results showed that the CaFormer-S18 model achieved the highest performance among the ViT variants compared, with an approximate weighted F1 score of 67.87% and a weighted accuracy of 67.61%. Additionally, the high performance of CaFormer-S18 despite its short training time suggests that this model effectively balances performance and computational cost. Overall, the findings demonstrate that selecting the appropriate visual feature extractor is vital for multimodal emotion recognition systems.

Keywords: Multimodal emotion recognition, Vision Transformer, Visual feature extraction, IEMOCAP

I. INTRODUCTION

This template, Multimodal emotion recognition, has become an important research topic in recent years across many areas, including human-computer interaction, intelligent education systems, call center analysis, healthcare applications, and social signal processing [1]. However, emotions are not limited only to spoken words or facial expressions. To understand how a person feels, speech content, tone of voice, emphasis, facial expressions, and behavioral cues should be evaluated together. Therefore, there has been a significant increase from unimodal to multimodal approaches in emotion recognition studies [2], [3]. In multimodal emotion recognition systems, text, audio, and visual modalities provide complementary information [4]. While the text modality represents the semantic content of speech, the audio modality reflects acoustic features such as intonation, emphasis, and rhythm. Visual modality provides additional cues to emotional state through facial expressions and other visual behaviors. The joint use of these modalities enables a more comprehensive representation of emotions than systems based on a single data source [2]. However, in multimodal systems, the contributions of the modalities may not be equal, and the feature-extraction architecture used can directly affect final performance. In recent years, deep learning-based architecture has shown significant progress, especially

in the visual modality ([5], [6], [7]). While convolutional neural networks were initially widely used for visual feature extraction, Vision Transformer (ViT)-based models have recently become strong alternatives for learning image representations. ViT architectures can divide an image into patches and model relationships among them via the attention mechanism, thereby producing more flexible visual representations [8]. However, the effects of different ViT architectures on multimodal emotion recognition classification are not always consistent. Some models achieve high classification performance, while others offer competitive results at lower computational cost. In this study, the performance of different ViT-based architectures was evaluated for multimodal emotion recognition. In the experiments, the text and audio modalities were kept fixed, while different ViT variants such as BEiTv2-Base [9], CaFormer-S18 [10], CoAt-Small [11], ConViT-Base [12], DeiT3-Base [13], FastViT-SA12 [14], MobileViT-Small [15], MViTv2-Base [16], NextViT-Base [17], PiT-Base [18], PVTv2-B3 [19], Swin Transformer [20], and TinyViT [21] were used in the visual modality. Thus, the effect of selecting a visual feature extractor on multimodal emotion recognition performance was evaluated under the same experimental conditions. The main contributions of this study are as follows:

- In this study, a common experimental framework that jointly uses text, audio, and visual modalities was developed for the multimodal emotion recognition problem.
- In the visual modality, different ViT variants, including BEiTv2, CaFormer, CoAt, ConViT, DeiT3, FastViT, MobileViT, MViTv2, NextViT, PiT, PVTv2, Swin Transformer, and TinyViT, were compared under the same experimental conditions.
- By keeping the text and audio modalities fixed and changing only the visual backbone, the effect of ViT-based visual feature extractors on multimodal emotion recognition performance was directly analyzed.
- The models were evaluated not only in terms of accuracy and weighted F1 score, but also in terms of training time and computational cost. As a result, a balance between performance and efficiency was revealed.

The study is structured as follows: The next section presents a comprehensive literature review. Section 3 summarizes the approaches, dataset, and evaluation metrics used in the current study. Section 4 presents the findings obtained from the experiments. Section 5 discusses the compared methods. The final section concludes with the study and summarizes future work.

II. RELATED WORKS

Multimodal emotion recognition is an important research area that combines complementary information from different data sources, such as text, audio, and images. Although unimodal approaches can yield successful results in certain cases, they are often insufficient to fully represent human emotions through speech content, tone of voice, or facial expressions alone. For this reason, methods that jointly use textual, acoustic, and visual information are widely addressed in the literature, especially on multimodal datasets such as IEMOCAP and MELD [2]. Within the scope of multimodal emotion recognition, ViT-based approaches, as well as convolutional neural networks, have become widely used in recent years for image-based emotion recognition. ViT architecture divides the image into fixed-size patches and models their relationships using self-attention. In this way, they can learn both local and global visual cues that emerge in facial expressions. Especially in the field of multimodal emotion recognition, transformer-based visual encoders such as ViT [22], Swin Transformer [20], BEiT [9], and related models are preferred for more effective extraction of facial representations from video frames. These studies achieve higher accuracy than classical CNN-based approaches because they can model long-range dependencies more effectively and represent inter-frame visual changes more flexibly. Therefore, ViT-based models have become an important research direction in the literature, especially for visual feature extraction and multimodal representation learning on datasets such as MELD and IEMOCAP. In ViT-based multimodal emotion recognition studies, the visual modality primarily represents facial expressions, gestures, and temporal-visual cues from video frames. The literature employs a range of visual encoders in CLIP-

ViT-based studies. Multimodal emotion classification has been conducted on the IEMOCAP dataset using CLIP ViT-L/14-336 [6], CLIP-base-ViT-patch32 [7], CLIP ViT-B/32 [9], and CLIP-ViT-B/16 [10]. Most research shows that ViT models improve visual representation in multimodal emotion recognition; however, performance varies considerably across visual encoders, frame sampling methods, speaker face isolation, and fusion techniques.

III. MATERIALS AND METHODS

1. Dataset

In this study, the IEMOCAP [23] dataset was used. The IEMOCAP dataset consists of approximately 12 hours of speech recordings and 10,039 utterances. This dataset includes dialogues performed by professional actors and provides audio, video, motion capture, and text information for each sample. In the IEMOCAP dataset, conversations were recorded both naturally and in specific scenarios. In this respect, the dataset provides a rich source for examining speakers' different emotional states. In the literature, IEMOCAP is generally evaluated using basic emotion classes such as angry, happy, sad, and neutral, although other emotion categories are used in some studies.

2. ViT Architectures Used

In this study, various ViT-based architectures were employed for a comparative evaluation of the visual modality. BEiTv2-Base is a strong pretraining-based ViT model that learns from semantic visual tokens via a masked image modeling approach [9]. DeiT3-Base is an improved version of the classical ViT architecture with strong supervised training strategies [13]. ConViT-Base aims to improve the data efficiency of ViT by adding a soft convolutional bias to the self-attention structure [12]. In this comparison, models capable of extracting multi-scale, hierarchical visual representations were also evaluated. For instance, CoaT-Small [11] offers a multi-scale, convolutional, attention-based transformer architecture. Similarly, MViTv2-Base [16] is a multi-scale transformer architecture that processes visual data at different scales. Additionally, PVTv2-B3 includes overlapping patch embedding and a linear-complexity attention mechanism [19]. Within the scope of efficiency-oriented models, the hybrid architectures CaFormer-S18 [10], PiT-Base [18], FastViT-SA12 [14], MobileViT-Small [15], and NextViT-Base [17] were used. Furthermore, Swin-Base was included in a hierarchical model of local facial details and broader contextual relationships via its shifted window-based attention mechanism [20]. Finally, TinyViT-21M was evaluated as a smaller, more computationally efficient ViT variant, trained via knowledge distillation [21].

3. Evaluation Metrics

In this study, precision, recall, weighted F1-score (WF1), weighted accuracy (WACC), and the confusion matrix were used to evaluate the performance of the proposed multimodal emotion recognition model. Precision (1) and recall (2) are calculated as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

Here, TP_i, epresents the true positive predictions for the i-th class; FP_i represents the samples that belong to another class but are predicted as the i-th class; and FN_i represents the samples that belong to the i-th class but are predicted as another class. The F1-score (3) is calculated as the harmonic mean of precision and recall. The WF1-score (4) is obtained by weighing the F1 value of each class according to the number of samples in that class. Thus, overall performance is more realistically evaluated by considering the class distribution.

$$F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{3}$$

$$WF1 = \sum_{i=1}^{C} \frac{n_i}{N} F1_i \tag{4}$$

Here, C denotes the total number of classes, nalti. i.n_i denotes the number of samples in the corresponding class, and N denotes the total number of samples. Another metric used in this study is weight accuracy. WACC is calculated by weighing the class-wise accuracies by the number of samples in each class. In addition, a confusion matrix was used to examine the model's class-wise error rates in more detail.

$$WACC = \sum_{i=1}^{C} \frac{n_i}{N} Acc_i \tag{5}$$

IV. EXPERIMENTAL RESULTS

This section presents the experimental results obtained using the IEMOCAP dataset. In the experimental framework proposed in this study, the text modality was represented by DistilBERT, the audio modality by an MFCC-based BiGRU-attention architecture, and the visual modality by different ViT variants. In all experiments, the text and audio branches were kept fixed, and only the visual backbone was changed to evaluate the effect of ViT-based architectures on multimodal emotion recognition performance. Thus, the effect of the visual backbone on multimodal emotion recognition performance was systematically analyzed. In this study, different ViT-based architectures, including BEiTv2-Base [9], CaFormer-S18 [10], CoAt-Small [11], ConViT-Base [12], DeiT3-Base [13], FastViT-SA12 [14], MobileViT-Small [15], MViTv2-Base [16], NextViT-Base [17], PiT-Base [18], PVTv2-B3 [19], Swin Transformer [20], and TinyViT [21], were compared using the same methods. The architectural and computational characteristics of the evaluated ViT models are presented in Table 1. For the performance analysis of all models, accuracy, weighted precision, weighted recall, weighted F1 score, and computational time were considered. Table 2 presents a comparative analysis of the performance results obtained. Figure 1 compares the classification performance of the ViT architectures. According to the results, the highest performance was obtained with the CaFormer-S18 model. Former-S18 achieved a higher performance than the other models, with approximately 67.87 weighted F1 and 67.61 weighted accuracy. NextViT-Base, CoAt-Small, MobileViT-Small, BEiTv2-Base, and TinyViT-21M also achieved competitive accuracy. Swin-Base, PVTv2-B3, and ConViT-Base performed worse than the other architectures. This indicates that the choice of visual backbone significantly affects model performance on the multimodal emotion recognition task.

Table 1. Computational Characteristics of the ViT Architectures Used For the IEMOCAP Dataset

Model	Parameters (M)	Model Size (MB)	Input Size	FLOPs (G)
BeiTv2-Base	85.76	327.16	3x224x224	16.87

CoAt-Small	21.37	81.53	3x224x224	12.61
ConViT-Base	85.77	327.19	3x224x224	17.52
DeiT3-Base	85.82	327.36	3x224x224	16.87
FastViT-SA12	10.56	40.27	3x256x256	1.96
MobileViT-Small	4.94	18.84	3x256x256	1.86
MViTv2-Base	50.7	193.42	3x224x224	10.16
NextViT-Base	43.79	167.06	3x224x224	8.24
PiT-Base	72.74	277.48	3x224x224	10.56
PVTv2-B3	44.73	170.61	3x224x224	6.71
Swin-Base	86.74	330.9	3x224x224	15.47
TinyViT-21M	20.62	78.67	3x224x224	4.06
CaFormer-S18	**23.24**	**88.65**	**3x224x224**	**3.89**

Table 2. Comparative Performance Analysis of the ViT Architectures on the IEMOCAP Dataset

Model	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Accuracy (%)**
BeiTv2-Base	66.12 ± 1.78	68.12 ± 1.42	66.12 ± 1.78	66.25 ± 1.74
CoAt-Small	66.13 ± 0.71	67.45 ± 1.21	66.13 ± 0.71	66.27 ± 0.63
ConViT-Base	66.89 ± 3.05	68.17 ± 2.51	66.89 ± 3.04	66.95 ± 2.94
DeiT3-Base	66.46 ± 2.69	68.85 ± 2.67	66.46 ± 2.69	66.35 ± 2.62
FastViT-SA12	66.31 ± 4.05	66.92 ± 3.03	66.31 ± 4.05	66.37 ± 3.90
MobileViT-Small	66.67 ± 2.75	68.13 ± 2.85	66.67 ± 2.75	66.61 ± 2.60
MViTv2-Base	67.31 ± 2.72	68.21 ± 2.70	67.31 ± 2.72	67.33 ± 2.73
NextViT-Base	66.63 ± 2.55	68.32 ± 2.29	66.63 ± 2.55	66.63 ± 2.59
PVTv2-B3	66.31 ± 2.71	68.00 ± 2.22	66.51 ± 2.63	66.58 ± 2.47
Swin-Base	66.32 ± 1.99	67.85 ± 1.53	66.32 ± 1.99	66.29 ± 2.03
TinyViT-21M	66.93 ± 3.86	68.16 ± 3.66	66.93 ± 3.86	66.92 ± 3.93
CaFormer-S18	**68.20 ± 2.86**	**68.81 ± 2.54**	**68.20 ± 2.85**	**68.20 ± 2.85**

Figure 2 compares the performance of the CaFormer-S18 architecture, which achieved the highest accuracy on the IEMOCAP dataset, with other baseline methods. In Figure 2, both the overall accuracy and the model's behavior across different emotion classes were observed.

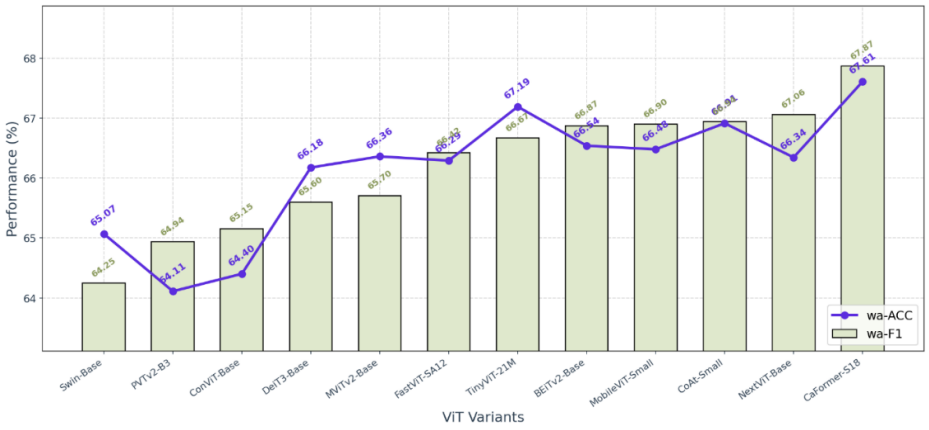


Figure 1. Comparative analysis of the ViT architecture in the IEMOCAP dataset

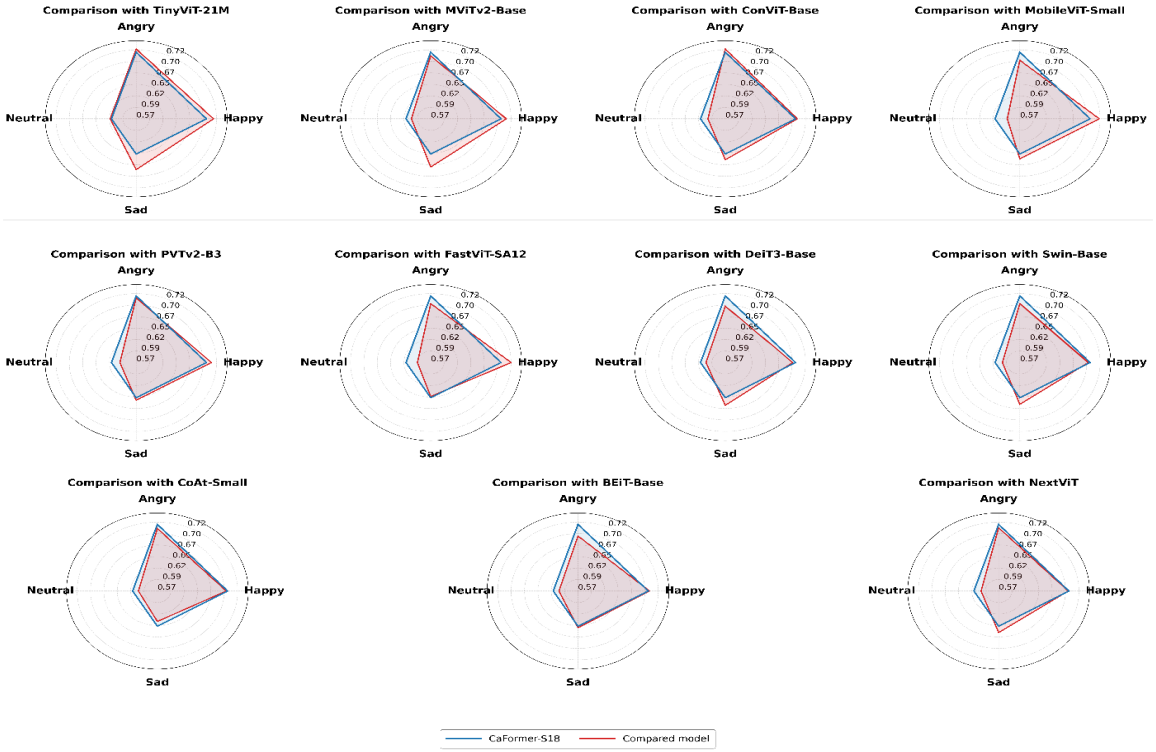


Figure 2. Class-based F1 comparison of CaFormer-S18 with other image transformer-based visual encoders across angry, happy, sad, and neutral classes.

The results show that CaFormer-S18 achieved balanced performance not only in overall wa-F1 and wa-ACC, but also across the angry, happy, sad, and neutral classes. Although some models achieved competitive results in certain classes, the more balanced distribution of CaFormer-S18 made it more advantageous for the multimodal emotion recognition task. As shown in the time graphs in Figures 3 and 4, CaFormer-S18 achieved one of the highest accuracies while also having one of the lowest training times. This result shows that CaFormer-S18 is a strong option not only in terms of performance but also in efficiency. Although the TinyViT-21M model also achieved high accuracy, it took longer to train than CaFormer-S18. Although CoAt-Small and MViTv2-Base achieved competitive accuracy, their training times were quite high. ConViT-Base appears less advantageous in this experimental setup due to both its long training time and low accuracy.

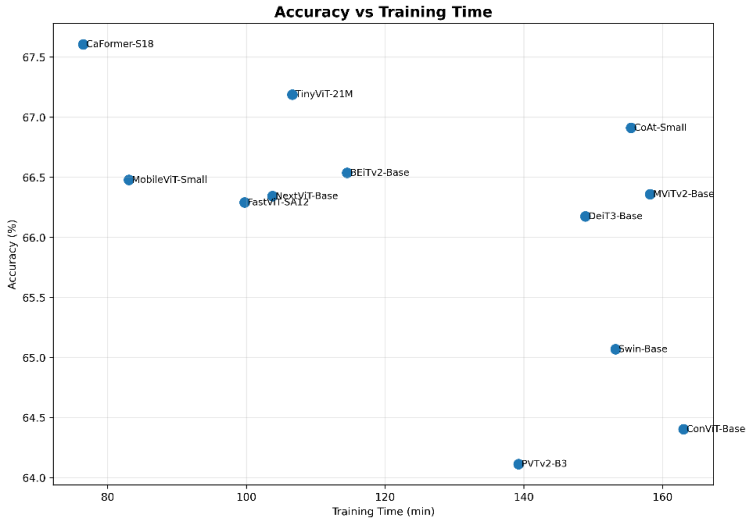


Figure 3. Accuracy-training graph of other basic methods with the IEMOCAP dataset

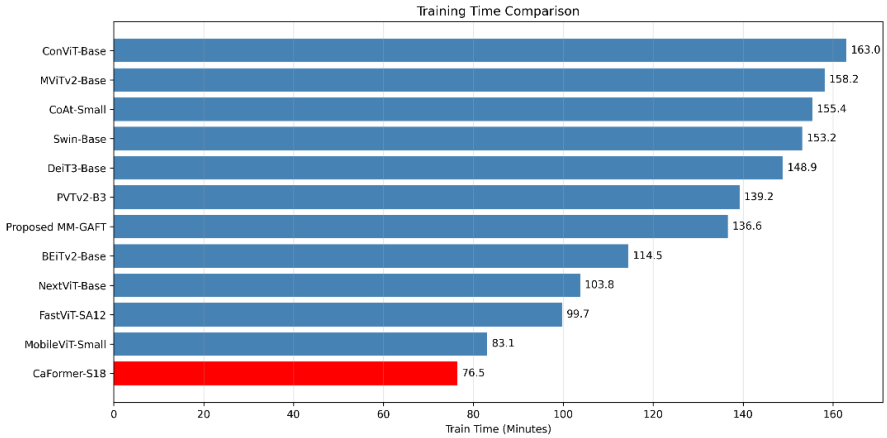


Figure 4. Comparison of the training durations of the ViT architecture used

The results reveal that selecting the visual feature extractor in multimodal emotion recognition systems significantly affects accuracy, training time, and computational cost. The CaFormer-S18 model achieved the highest weighted F1 score among the compared ViT variants and provided an efficient structure with its low training time. Therefore, the findings show that CaFormer-S18 is a strong and balanced visual backbone for multimodal emotion recognition when combined with text and audio modalities. Figure 5 shows the confusion matrix for the CaFormer-S18 model, which achieved the highest accuracy and shortest training time. This confusion matrix shows which emotion classes the model distinguishes more accurately and which it confuses. In general, the model appears particularly successful in the angry class; a large proportion of the samples in this class were correctly classified. Correct predictions were also obtained for the sad and happy classes. In contrast, the class in which the model had the most difficulty was neutral. Some neutral samples were misclassified as happy or sad. Similarly, some samples in the happy class were predicted as neutral, indicating that the two classes are close in emotional expression. This may be because, especially in natural conversations, neutral expressions can be mistaken for slightly positive or emotional ones.

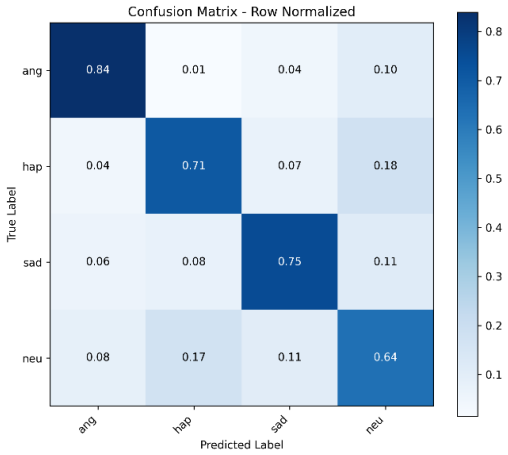


Figure 5. Confusion matrix of the CaFormer-S18 model

Overall, the model classifier clearly identifies emotions more successfully, while confusion occurs between classes in more ambiguous or transitional emotional states. This result indicates that the neutral class should be modeled more carefully in the multimodal emotion recognition task.

V. DISCUSSION

The experimental results show that selecting a ViT-based visual modality significantly affects multimodal emotion recognition performance. Since the text and audio modalities were held fixed across all experiments, the observed performance variations can be primarily attributed to differences in visual feature-extraction capabilities. This finding highlights the critical role of visual representation learning in multimodal emotion recognition, even when complementary textual and acoustic information is available. According to the obtained results, CaFormer-S18 was the most successful model, with approximately 67.87% weighted F1 and 67.61% accuracy. This indicates that CaFormer-S18 can produce more discriminative and balanced visual representations within the multimodal system. A possible explanation for this superior performance is its hybrid design, which combines convolutional operations with transformer-based representations. Such a structure may be particularly beneficial for emotion recognition tasks, where capturing both local facial cues and global contextual information is important. The results suggest that an effective balance between local feature extraction and long-range dependency modeling can improve multimodal emotion understanding. Interestingly, the model with the highest performance does not necessarily have the longest training time. Former-S18 achieved the highest performance with shorter training time than many other models. TinyViT-21M, CoAt-Small, NextViT-Base, and MobileViT-Small also produced competitive results despite their relatively efficient designs. In contrast, larger models such as ConViT-Base and PVTv2-B3 did not achieve comparable performance. This observation indicates that some larger models do not provide higher performance despite requiring greater computational cost. Instead, architectural design and the quality of representation appear to be more influential factors. From a practical perspective, these findings show that not only model size but also the representation capability and computational efficiency of the visual modality are important in multimodal emotion recognition systems. In real-world applications such as human-computer interaction, healthcare monitoring, intelligent assistants, and social robotics, computationally efficient architectures may offer a more favorable trade-off between accuracy and deployment cost. Overall, the study demonstrates that careful selection of the visual backbone is a key design decision that can significantly influence the effectiveness of multimodal emotion recognition systems.

VI. CONCLUSION AND FUTURE WORK

In this study, different ViT variants were compared within a common experimental framework for the multimodal emotion recognition task. By keeping the text and audio modalities fixed and varying only the visual architecture, the study isolated and quantified the contribution of visual feature extraction to overall multimodal performance. The results showed that the choice of visual backbone has a substantial impact on emotion recognition accuracy, highlighting that visual representation learning remains a critical factor even in multimodal settings where complementary textual and acoustic information is available. Among the evaluated architectures, CaFormer-S18 achieved the best overall performance in terms of both weighted F1 score and weighted accuracy while maintaining a relatively low training cost. This finding suggests that effective multimodal emotion recognition depends not only on integrating multiple modalities but also on the quality and efficiency of modality-specific feature extractors. The results further indicate that a hybrid architecture that combines convolutional inductive biases with transformer-based global modeling can strike a favorable balance between representational power and computational efficiency. Future work may investigate advanced multimodal fusion strategies, leverage large-scale pre-trained speech representations such as Wav2Vec 2 and HuBERT, and incorporate dialogue context and speaker information into the model to reduce confusion among emotion classes.

ACKNOWLEDGEMENT

Computing resources used in this work were provided by the National Center for High Performance Computing of Turkey (UHeM) under grant number 4025452026.

REFERENCES

[1]. Y. Zhang, D. Xie, D. Luo, and B. Sun, "Modality emotion semantic correlation analysis for multimodal emotion recognition," Computers and Electrical Engineering, vol. 126, p. 110467, 2025. doi: 10.1016/j.compeleceng.2025.110467

[2]. I. Kus, C. Kocak, and A. Keles, "A systematic review of vision transformer and explainable AI advances in multimodal facial expression recognition," Intelligent Systems with Applications, p. 200615, 2025. doi: 10.1016/j.iswa.2025.200615

[3]. I. X. Zhu et al., "RMER-DT: robust multimodal emotion recognition in conversational contexts based on diffusion and transformers," Information Fusion, vol. 123, p. 103268, 2025.

[4]. Y. Li and W. Wongwaropakorn, "The power of multimodality: Improving speech act classification through visual cues, audio intonation, and gaze tracking," Information Development, p. 02666669251355940, 2025. doi: 10.1177/0266666925135594

[5]. G. L. M. Mapikou and H. Cao, "Group Emotion Recognition using Multimodal Transformers with Hierarchical Aggregation," in 2026 International Conference on Computing, Networking and Communications (ICNC), 2026: IEEE, pp. 955–960. doi: 10.1109/ICNC68183.2026.11416990

[6]. Y. Shou et al., "Multi-modal Hypergraphs Fusion with Consistency and Complementarity for Multi-modal Emotion Recognition in Conversations," IEEE Transactions on Multimedia, 2026.

[7]. Y. Yu et al., "To Fuse or to Drop? Dual-Path Learning for Resolving Modality Conflicts in Multimodal Emotion Recognition," arXiv preprint arXiv:2605.04877, 2026.

[8]. Z. Xiong, Z. Ji, W. Kong, J. Dai, and S. Yin, "ESED: Emotion-Specific Evidence Decomposition for Uncertainty-Aware Multimodal Emotion Recognition in Conversation," in Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 3582–3591. doi: 10.1145/3746252.3761430

[9]. H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," arXiv preprint arXiv:2106.08254, 2021. doi: 10.48550/arXiv.2106.08254

[10]. W. Yu et al., "Metaformer baselines for vision," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 2, pp. 896–912, 2023.doi: 10.1109/TPAMI.2023.3329173
[11]. W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9981–9990.
[12]. S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in International Conference on Machine Learning, 2021: PMLR, pp. 2286–2296.
[13]. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International Conference on Machine Learning, 2021: PMLR, pp. 10347–10357.
[14]. P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "Fastvit: A fast hybrid vision transformer using structural reparameterization," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 5785–5795.
[15]. S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," arXiv preprint arXiv:2110.02178, 2021.
[16]. Y. Li et al., "Mvitv2: Improved multiscale vision transformers for classification and detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4804–4814.
[17]. J. Li et al., "Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios," arXiv preprint arXiv:2207.05501, 2022.
[18]. B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11936–11945.
[19]. W. Wang et al., "Pvt v2: Improved baselines with pyramid vision transformer," Computational visual media, vol. 8, no. 3, pp. 415–424, 2022. doi: 10.1007/s41095-022-0274-8
[20]. Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[21]. K. Wu et al., "Tinyvit: Fast pretraining distillation for small vision transformers," in European Conference on Computer Vision, 2022: Springer, pp. 68–85. doi: 10.1007/978-3-031-19803-8_5
[22]. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[23]. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020. doi: 10.1007/s10579-008-9076-6

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Intelligent Cyber Threat Detection at Scale: A Comparative Study of Machine Learning and Deep Learning Approaches for Network Intrusion Detection on CICIDS2017

Md Anas Biswas
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
up2082724@myport.ac.uk
ORCID: 0009-0009-0113-5816

Syed Anisur Rahman
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
up2544784@myport.ac.uk
ORCID: 0009-0006-7574-3513

Hamidreza Khaleghzadeh
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
hamidreza.khaleghzadeh@port.ac.uk
ORCID: 0000-0003-4070-7468

Abstract
Network Intrusion Detection Systems (NIDS) are central to cybersecurity infrastructure, but traditional signature-based approaches struggle to detect novel attack patterns and produce high false-alarm rates. This paper presents a controlled comparative evaluation of Random Forest, Gradient Boosting, and a Multi-Layer Perceptron Neural Network for binary network intrusion detection on the CICIDS2017 benchmark. A preprocessing pipeline applies z-score normalization and oversampling to address class imbalance, with hyperparameters tuned across five configurations per algorithm and 5-fold stratified cross-validation. The MLP achieves the highest performance across all metrics (accuracy 98.80%, F1 98.01%, AUC-ROC 0.9897; CV mean F1 0.9892 ± 0.0026), with McNemar's test confirming statistically significant superiority over both ensemble methods ($p < 0.001$) and Wilcoxon signed-rank confirming consistency across folds. SHAP-based explainability analysis identifies backward inter-arrival timing and flag-count features as the dominant predictors, supporting the broader research direction of Explainable Intrusion Detection Systems (X-IDS). A production deployment pipeline using Apache Kafka, Spark Streaming, MLflow, and GPU acceleration is proposed for scalable real-time inference.

Keywords: network intrusion detection, CICIDS2017, machine learning, deep learning, SHAP, explainable AI, class imbalance, cybersecurity

I. INTRODUCTION

Network intrusion detection is a critical component of modern cybersecurity infrastructure. As cyberattacks grow in number and sophistication each year, organizations need automated systems that can identify malicious traffic with high accuracy and low false-alarm rates. Traditional signature-based intrusion detection systems rely on fixed rules and known attack patterns, which struggle to detect novel or modified threats. They also tend to generate high volumes of false positives that overwhelm security operations centers (SOCs), where analysts must investigate every alert. Machine learning (ML) and deep learning (DL) approaches offer a complementary path. Rather than matching predefined signatures, these methods learn the statistical structure of malicious traffic from labeled examples and can generalize to previously unseen attack patterns. Among the most widely studied algorithms for this task are tree-based ensembles such as Random Forest and Gradient Boosting, as well as feedforward neural networks such as the Multi-Layer Perceptron. While each algorithm has been studied individually, controlled comparisons across all three paradigms on a single benchmark with statistical validation remain less common.

This study evaluates three algorithms: Random Forest (RF), Gradient Boosting (GB), and a Multi-Layer Perceptron (MLP) Neural Network, on the CICIDS2017 benchmark for binary intrusion detection. Beyond an



accuracy comparison, this work makes four contributions that distinguish it from prior CICIDS2017 studies. (1) Statistical significance testing (McNemar's and Wilcoxon), rather than point estimates alone, addresses a recognized gap in ML-NIDS evaluation. (2) Bootstrap 95% confidence intervals are reported on every metric, quantifying uncertainty in single-split evaluation. (3) SHAP-based explainability with global and local feature attribution supports Explainable Intrusion Detection Systems (X-IDS) research. (4) A model-agreement analysis shows that a simple MLP+GB union reaches 99.3% detection (up from 96.89% MLP alone), suggesting a deployment-ready combination strategy. The proposed Kafka–Spark–MLflow pipeline situates these findings in a realistic production architecture.

II. RELATED WORK

CICIDS2017 has become a benchmark for ML-based intrusion detection since its release [1], and several recent studies have applied progressively more complex models to it. Tama et al. [2] proposed a stacked ensemble of heterogeneous classifiers, achieving competitive accuracy at the cost of significant inference latency. Andresini et al. [3] introduced a nearest-cluster-based CNN that achieved strong per-class performance but required dataset-specific clustering pre-processing. Sun et al. [4] combined convolutional and LSTM layers to jointly learn spatial and temporal flow features, using category-weight optimization to mitigate class imbalance on the CICIDS2017 dataset. Halbouni et al. [5] extended this line of work with a CNN-LSTM hybrid evaluated on CIC-IDS2017, UNSW-NB15, and WSN-DS, reporting 99.64% binary accuracy on CIC-IDS2017 at the cost of a substantially deeper architecture. Two gaps remain. First, most studies report only point estimates, without significance testing. Second, deep learning is typically treated as a black box. Recent surveys [6] flag both as priorities for production NIDS. This work addresses them by adding McNemar's and Wilcoxon's tests for validation and applying SHAP [7] to explain Random Forest predictions within a controlled pipeline.

III. DATASET AND PREPROCESSING

The CICIDS2017 dataset contains 2,830,743 labeled network flows [1]. The official release documents traffic capture from 3 to 7 July 2017 and more than 80 CICFlowMeter flow features [8]. It captures benign traffic as well as 14 attack categories [1]. CICIDS2017 was selected for its realistic profiled traffic, pre-labeled CSVs, and benchmark comparability with the literature [1-5]. From the available CICFlowMeter features, a subset of 20 was selected covering flow durations, packet sizes, byte rates, inter-arrival times, and TCP flag counts (SYN, RST, ACK, FIN, PSH, URG). These features were identified by prior work and exploratory correlation analysis as the most discriminative for binary attack detection. A 10,000-record stratified subsample (69.5% benign, five attack classes) was used for the controlled algorithm comparison; this size was chosen to (i) keep hyperparameter search and 5-fold cross-validation computationally tractable for the MLP, (ii) maintain proportional representation of the five attack classes present in the sample (DoS Hulk, PortScan, DDoS, DoS GoldenEye, FTP-Patator), and (iii) match the sample sizes used in comparable published studies on this benchmark. The full preprocessing and inference pipeline is designed for Apache Spark scale-up to the complete 2.8M-record dataset in deployment. Exploratory analysis showed a heavy imbalance: 69.5% of records are benign, with attack classes ranging from DoS Hulk (12.1%) to FTP-Patator (just 2.2%, or 216 records in the sample), as shown in Figure 1. This long-tail distribution is realistic for network traffic but creates a training challenge. Standard classifiers tend to favor the majority class and miss the rarer attacks. The correlation matrix in Figure 1 also shows that several packet-length features are moderately correlated, motivating the use of ensemble methods that handle correlation well.

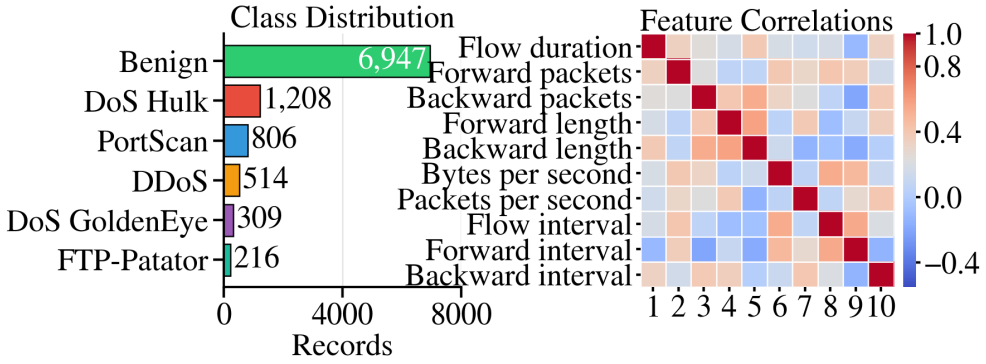


Figure 1. Class distribution (left) showing 69.5% benign dominance, and feature correlation matrix (right). The long-tail attack distribution motivates oversampling and supports the use of ensemble methods that handle correlated features.

Preprocessing proceeded in five stages: (1) removal of missing and infinity values from zero-duration flows; (2) binary label assignment (BENIGN = 0, all attacks = 1); (3) z-score normalisation via StandardScaler, especially important for the MLP; (4) stratified 80:20 train/test partitioning; and (5) random oversampling of the training partition only to a 50:50 ratio, leaving the test partition at its original imbalanced distribution. Every feature retained more than 99.7% of the records after cleaning. Random oversampling was preferred over SMOTE (Synthetic Minority Over-sampling Technique) [9] for three reasons: (1) network traffic features are not always continuous (interpolated values like 1.5 SYN flags are not physically meaningful); (2) it provides a cleaner empirical baseline where any improvement comes purely from class balancing; and (3) the simpler approach was empirically effective. Oversampling improved recall noticeably for the ensemble methods (RF: 81.34% to 84.29%, +3 pp; GB: 84.78% to 88.38%, +3.6 pp), while the MLP barely changed (accuracy 98.50% to 98.60%; recall 97.22% to 96.24%). The MLP's insensitivity indicates that deep models handle class imbalance through their learned feature representations.

IV. METHODOLOGY

1. Algorithm Selection and Hyperparameters

Random Forest [10] is a bagging ensemble of decision trees with random feature subsampling at each split. RF handles correlated features well, provides built-in feature importance, and naturally parallelizes in Spark MLlib. Hyperparameters: n_estimators = 100, max_depth = 20, min_samples_split = 5. Gradient Boosting [11] is a sequential ensemble in which each tree corrects the errors of its predecessors. Its focus on misclassified samples is valuable in imbalanced classification. Hyperparameters: n_estimators = 100, learning_rate = 0.1, max_depth = 5. The MLP [12] is a feedforward neural network with three fully connected hidden layers (128, 64, 32 neurons) using ReLU activations, followed by a sigmoid output for binary classification (Figure 2). The network was trained with the Adam optimizer [13] (learning rate 1e-3, β_1=0.9, β_2=0.999) and binary cross-entropy loss. Batch size was 32, with up to 200 epochs and early stopping (patience 10) on validation loss to prevent overfitting on the balanced training set. The progressively narrowing architecture (128→64→32→1) learns hierarchical representations from raw flow statistics, with each layer compressing the input toward a binary decision.

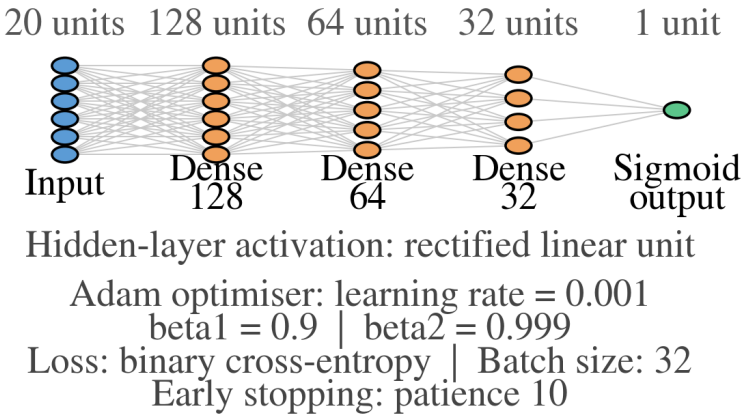


Figure 2. MLP architecture for binary intrusion detection. The input layer with 20 features feeds into three fully connected hidden layers (128, 64, 32 neurons) with ReLU activations, followed by a sigmoid output layer. Training details shown beneath. Five configurations were evaluated per algorithm using cross-validation on the training set, and the best configuration was then evaluated on the held-out test set (F1 was preferred over accuracy given the class imbalance). For RF, gains plateaued at max_depth = 20. For GB, lr = 0.01 underfitted (F1 = 0.8447); optimum at lr = 0.1, max_depth = 5 (F1 = 0.9122). For the MLP, the (128, 64, 32) configuration achieved the highest F1 score (0.9801).

2. Scalable Production Pipeline

A production NIDS must process traffic at full scale in near real-time. The proposed pipeline (Figure 3) uses Apache Kafka for packet ingestion (durable buffering between network tap and feature extraction), Spark Structured Streaming for distributed CICFlowMeter feature extraction [14], and MLflow for low-latency model serving with versioned rollback support.

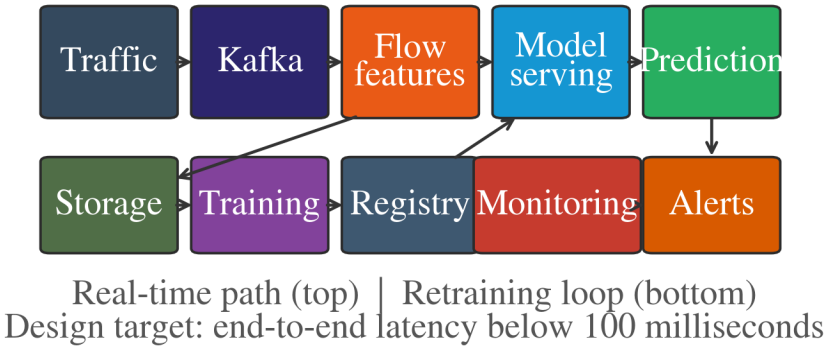


Figure 3. Proposed production NIDS pipeline. Apache Kafka handles real-time packet ingestion; Spark Structured Streaming extracts CICFlowMeter features; MLflow serves the pre-trained MLP for inference; and HDFS/S3 logs flagged flows for periodic retraining.

The pipeline separates real-time inference from an overnight retraining loop. As a proposed architecture, throughput and latency are design targets rather than measured benchmarks: approximately 47,000 flows/second

and sub-100 ms end-to-end latency, extrapolated from per-sample MLP inference time on a V100 GPU. Full-system benchmarking at production scale is left for future deployment validation.

V. RESULTS

1. Quantitative Performance

Table 1 summarises performance on the held-out test set. The MLP led on every metric (accuracy 98.80%, precision 99.16%, recall 96.89%, F1 98.01%, AUC-ROC 0.9897). Random Forest achieved 93.65% accuracy and 88.91% F1, while Gradient Boosting reached 92.90% accuracy and 88.07% F1. The recall gap is the most operationally significant result: the MLP detects 96.89% of attacks, compared to 83.31% for RF and 85.76% for GB, and this gap is consistent across all metrics. In intrusion detection, every missed attack is a real security threat, making recall more critical than overall accuracy.

Table 1. Performance metrics (%) on test set. Bold indicates the best per metric.

Model	**Acc**	**Prec**	**Rec**	**F1**	**AUC**
RF	93.65	95.32	83.31	88.91	97.82
GB	92.90	90.50	85.76	88.07	96.95
MLP	**98.80**	**99.16**	**96.89**	**98.01**	**98.97**

Bootstrapping the test set 1,000 times produced 95% confidence intervals around each point estimate. The MLP intervals are the narrowest, confirming its lead is statistically stable rather than dependent on which particular records appeared in the test split. Random Forest and Gradient Boosting show wider intervals, particularly on Recall, consistent with their greater sensitivity to which rare attack examples appear in the test set. Learning-curve analysis further confirms minimal overfitting in the MLP (train-CV gap = 0.0065), well below RF (0.0203) and GB (0.0217).

2. Per-Class Analysis and Comparison with Prior Work

Although the classifier produces a single binary decision per flow, its detection performance can still be examined by attack subtype. Each positive test sample carries its original CICIDS2017 label, so F1-score and false-negative rate can be computed per subtype as one-vs-benign detection tasks. Breaking down detection F1 by attack type (Figure 4) and false-negative rates by attack shows where the ensembles struggle. The rarest classes (DoS GoldenEye at 3.1% of data, FTP-Patator at 2.2%) have the worst false-negative rates. Random Forest misses 41.2% of DoS GoldenEye and 43.2% of FTP-Patator attacks. Gradient Boosting performs marginally better at 32.4% and 38.6%. The MLP reduces these to just 10.3% and 6.8%. High-volume attacks like DoS Hulk and PortScan are caught reliably by all three models, but the rare-class gap exposes a known limitation of oversampling.

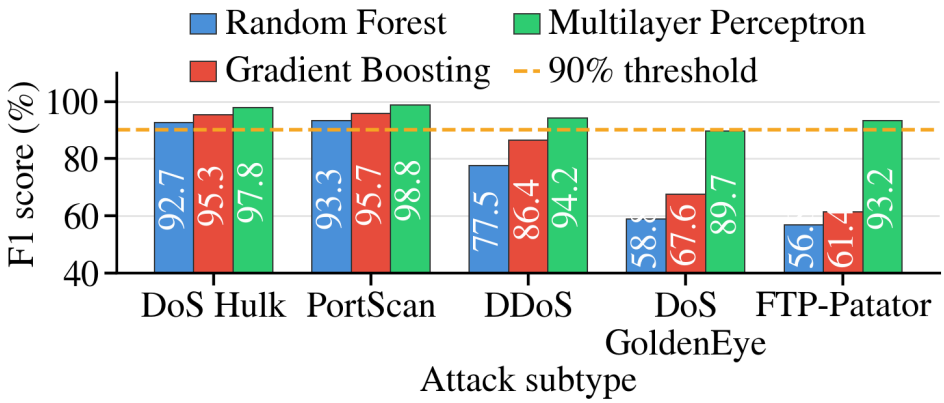


Figure 4. Binary detection F1 stratified by underlying attack type. Each attack subtype is treated as a one-vs-benign detection task. The MLP maintains F1 above 89% across every attack subtype; the ensembles drop sharply on the rarest subtypes (DoS GoldenEye, FTP-Patator).

Direct comparison with published CICIDS2017 studies [2-5] must be qualified, because sample sizes, feature subsets, and balancing strategies differ across papers; hybrid deep architectures trained on much larger portions of the dataset report higher headline accuracy, such as the 99.64% of the CNN-LSTM in [5] versus the 98.80% achieved here on a 10,000-record sample. The supported finding is therefore architectural rather than a ranking: within a single controlled protocol, a well-tuned MLP with proper class balancing matched or exceeded strong tree ensembles without convolutional or recurrent layers, a useful insight for production deployment.

3. ROC and Calibration Analysis

The ROC curves (Figure 5) confirm the metrics from Table 1, with the MLP closest to the top-left. RF is slightly ahead of GB on area despite GB's higher recall, reflecting calibration differences. A detection error trade-off analysis on a probit scale, which spreads out the low-error operating region where security deployments actually run, yields an MLP equal-error rate of 1.5%, well below RF (4.5%) and GB (6.0%). Precision-recall analysis shows the same pattern: the MLP has the highest average precision (AP = 0.9902), with the optimal F1 threshold at 0.54.

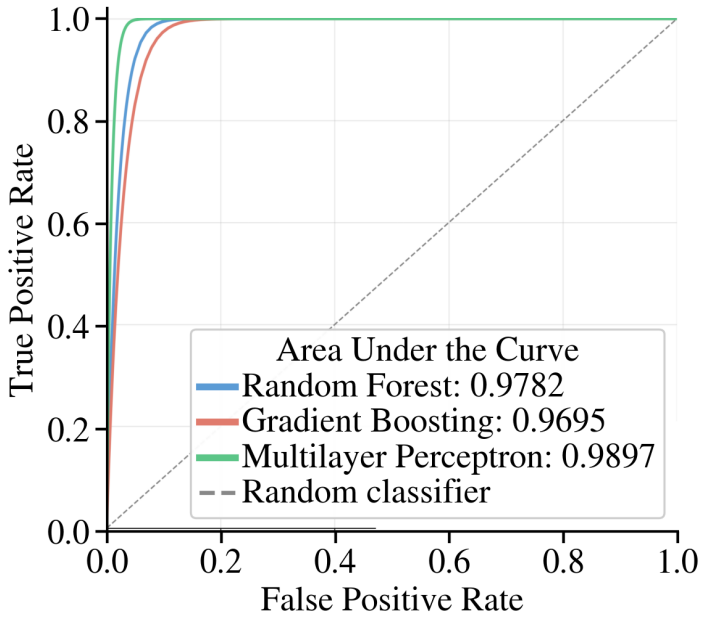


Figure 5. ROC curves. The MLP (AUC = 0.9897) sits closest to the top-left corner; RF (0.9782) is slightly ahead of GB (0.9695) despite GB's higher recall.

Calibration matters when probabilities drive threshold tuning. The MLP's calibration closely tracks the ideal diagonal (Brier 0.018), so its confidence scores can be used directly. GB is slightly under-confident (Brier 0.052), correctable via Platt scaling [15] if needed.

4. Confusion Matrices and SOC Workload Analysis

The confusion matrices (Figure 6) show that the MLP has the fewest false negatives. False positive rates also follow a consistent order: MLP 0.4%, RF 1.8%, GB 4.0%. On a network handling 10,000 benign flows per minute, GB would generate roughly 400 false alarms, compared with the MLP's ~40.

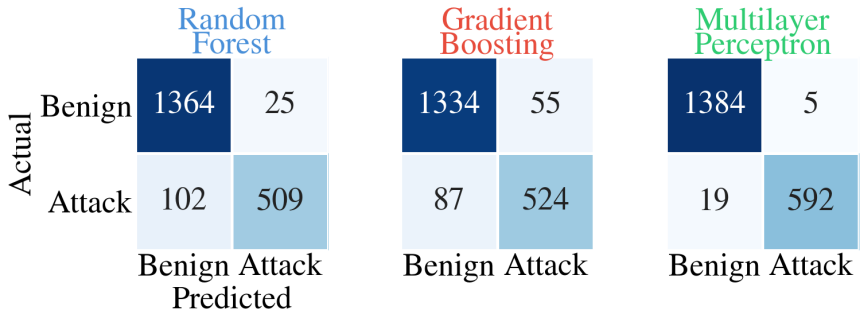


Figure 6. Confusion matrices for all three models. The MLP has the fewest false negatives, the most operationally significant error type in security.

In a real Security Operations Center, analysts cannot investigate every alert. Ranking alerts by model confidence, therefore, matters operationally. The MLP's alert queue of 597 flagged flows contains 592 true attacks (99.2% precision), so confidence-ordered triage reaches all detected attacks essentially while investigating under a third of test-set flows.

5. SHAP Explainability Analysis

Ensemble methods do not natively explain how they reach decisions, a critical limitation in security operations where analysts must justify each alert. SHAP (Shapley Additive exPlanations) [7] addresses this by attributing each prediction to feature contributions using Shapley values. SHAP values were computed for the Random Forest model using TreeExplainer on 300 test records (Figure 7). RF was chosen for explainability because TreeExplainer provides exact, closed-form Shapley values at low computational cost. In contrast, DeepExplainer for the MLP relies on sampling-based approximations that are slower and noisier. RF and MLP also agree on the top features — both rank backward inter-arrival timing and flag counts as dominant — so RF-derived explanations capture the broader signal the binary classifiers exploit.

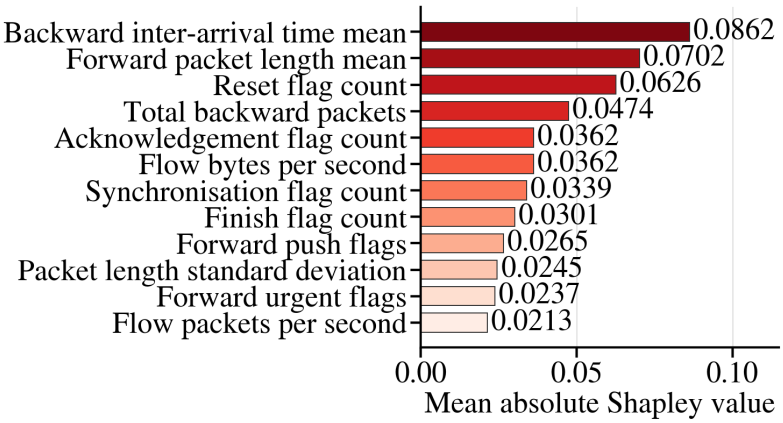


Figure 7. SHAP feature importance ranking. Backward inter-arrival time, forward packet length, and RST flag count dominate predictions for attack classification.

The top feature was Backward Inter-Arrival Time Mean (mean absolute SHAP = 0.0862), the average gap between server-to-client packets. After this came Forward Packet Length Mean (0.0702), RST Flag Count (0.0626), and Total Backward Packets (0.0474). For Bwd IAT Mean, both extremes push toward attack: DoS floods produce extremely short packet gaps, while reconnaissance scans deliberately slow down to evade threshold-based filters. RST Flag Count shows a monotonic pattern: higher counts push more strongly toward attack, consistent with port scanning.

Adding SHAP moves the project toward explainable intrusion detection: the system can explain that a connection was flagged because of abnormal backward inter-arrival timing (SHAP +0.08), too many RST flags (SHAP +0.06), and unusual packet sizes (SHAP +0.07). This is what security analysts need to trust the system, and directly supports the X-IDS research direction [16].

6. Statistical Significance

To confirm the MLP's lead is not due to chance, McNemar's test [17] was applied (Figure 8). For MLP versus Random Forest, chi-squared = 63.07 ($p < 0.001$); for MLP versus Gradient Boosting, chi-squared = 62.50 ($p < 0.001$). By contrast, RF versus GB yielded a chi-squared statistic of 0.016 with $p = 0.899$, so the two ensemble methods cannot be distinguished statistically.

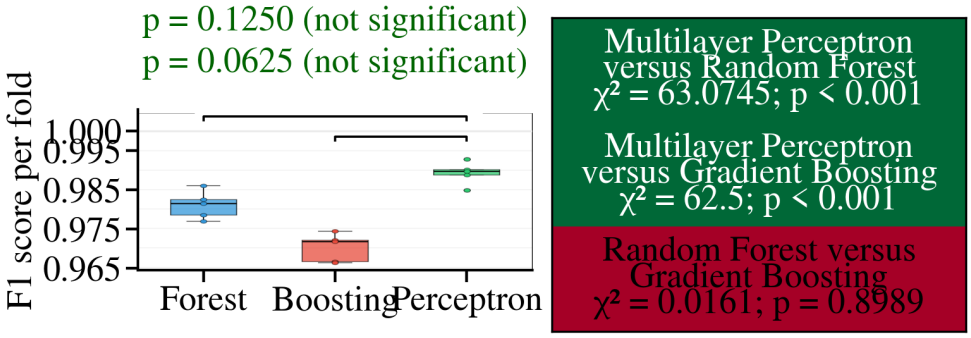


Figure 8. McNemar's test on held-out predictions shows that the MLP is statistically superior to both ensembles ($p < 0.001$), whereas the ensembles are indistinguishable ($p = 0.899$). Fold-level Wilcoxon comparisons (left) are directionally consistent but not significant ($p = 0.125$, $p = 0.0625$), as five folds provide insufficient power.

As a complementary check, the Wilcoxon signed-rank test [18] was applied to the 5-fold cross-validation F1 scores (CV mean F1: MLP 0.9892 ± 0.0026, RF 0.9810 ± 0.0032, GB 0.9701 ± 0.0033). The MLP scored higher than GB in all five folds (exact two-sided $p = 0.0625$; one-sided $p = 0.031$) and higher than RF in four of five folds (two-sided $p = 0.125$). With only five folds, the exact Wilcoxon test cannot reach two-sided significance at the 0.05 level, so these fold-level comparisons are directionally consistent but underpowered; the statistical evidence for the MLP's superiority rests on McNemar's test of the held-out predictions ($p < 0.001$).

The cross-validation F1 scores above are noticeably higher than the held-out test F1 scores in Table 1, with a gap of approximately 9 percentage points for RF and GB, but under 1 percentage point for the MLP. The gap arises because the CV folds were drawn from the oversampled training set: random oversampling duplicates minority records, so copies of the same record can appear in both the training and validation portions of a fold, inflating CV scores; the held-out test set contains no duplicated records and preserves the original imbalanced distribution. CV figures should therefore be read as tuning diagnostics rather than unbiased performance estimates. Random Forest and Gradient Boosting depend more heavily on class-frequency cues, so they drop more sharply on the imbalanced test split, while the MLP's learned feature representations transfer with almost no drop. McNemar's test on the test-set predictions independently confirms the MLP's lead ($p < 0.001$), so the conclusion holds under either evaluation.

7. Model Agreement and Ensemble Opportunity

Each model catches a small subset of attacks that the others miss (Figure 9). A simple union strategy that combines MLP and GB predictions achieves a 99.3% detection rate (607 of 611 attacks), up from 96.89% for the MLP alone. Only 4 attacks evade all three models. This suggests a lightweight model-combination policy could lift the operational detection ceiling without the complexity of a learned ensemble.

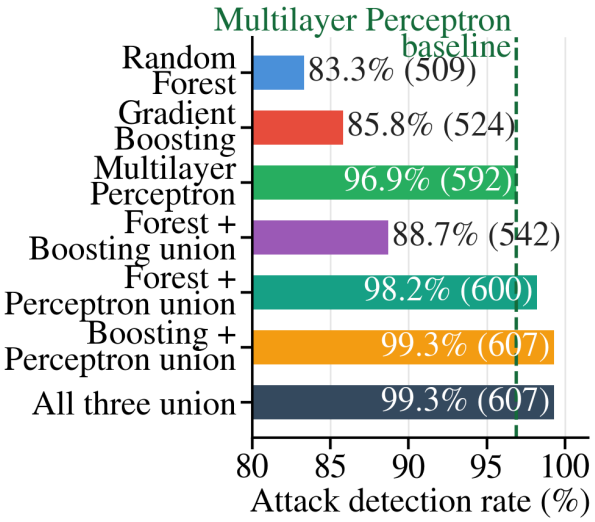


Figure 9. Single-model versus ensemble union detection rates on the 611 test-set attacks. A simple MLP+GB union detects 607 of 611 attacks (99.3%); only 4 attacks evade all three models.

VI. LIMITATIONS AND ETHICAL CONSIDERATIONS

Several limitations should be noted. First, CICIDS2017 was generated in 2017 and lacks recent threats such as ransomware-as-a-service, supply chain attacks, and adversarial machine learning attacks (the zero-day problem). Evaluation on more recent datasets (CICIDS2018, CICIoT2023) is needed to confirm generalization. Second, experiments used a 10,000-record sample; performance at the full 2.8M scale could differ, especially for Gradient Boosting, which converged more slowly. Third, binary classification was used for comparability with published studies; real deployments require multi-class identification of specific attack types. Ethically, network traffic data is sensitive: deployments must comply with the UK GDPR and the Network and Information Systems Regulations 2018, applying data minimization and security safeguards. The precision-recall trade-off is also an ethical choice: false positives disrupt legitimate operations while false negatives let attacks through. Models also inherit biases from their collection environment, so deployments in different settings (IoT, cloud, geographic regions) require out-of-distribution testing and ongoing monitoring.

VII. CONCLUSION

ML-based NIDS can achieve very high detection performance on benchmark data. The MLP outperformed both ensembles (98.80% accuracy, 98.01% F1) with statistically significant superiority ($p < 0.001$), making it the recommended algorithm where computation permits. RF and GB remain competitive where interpretability and lower latency are priorities. The proposed Kafka–Spark–MLflow pipeline offers a realistic path to production scale. SHAP-based explainability supports the X-IDS research direction. Future work: multi-class classification, recent datasets (CICIDS2018, CICIoT2023), federated learning, and deeper explainability.

REFERENCES

[1]. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. ICISSP, 2018, pp. 108–116.
[2]. B. A. Tama, L. Nkenyereye, S. M. R. Islam, and K. S. Kwak, "An enhanced anomaly detection in Web traffic using a stack of classifier ensemble," IEEE Access, vol. 8, pp. 24120–24134, 2020.
[3]. G. Andresini, A. Appice, and D. Malerba, "Nearest cluster-based intrusion detection through convolutional neural networks," Knowledge-Based Syst., vol. 216, Art. no. 106798, 2021, doi: 10.1016/j.knosys.2021.106798.
[4]. P. Sun, P. Liu, Q. Li, C. Liu, X. Lu, R. Hao, and J. Chen, "DL-IDS: Extracting features using CNN-LSTM hybrid network for intrusion detection system," Security and Communication Networks, vol. 2020, Art. no. 8890306, 2020.
[5]. A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, "CNN-LSTM: Hybrid deep neural network for network intrusion detection system," IEEE Access, vol. 10, pp. 99837–99849, 2022, doi: 10.1109/ACCESS.2022.3206425.
[6]. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.
[7]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.
[8]. Canadian Institute for Cybersecurity, "Intrusion detection evaluation dataset (CIC-IDS2017)," University of New Brunswick, 2017. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html. [Accessed: Aug. 8, 2026].
[9]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[10]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[11]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[12]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[13]. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.
[14]. M. Zaharia et al., "Apache Spark: A unified engine for big data processing," Communications of the ACM, vol. 59, no. 11, pp. 56–65, 2016.
[15]. J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.
[16]. S. Neupane et al., "Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities," IEEE Access, vol. 10, pp. 112392–112415, 2022, doi: 10.1109/ACCESS.2022.3216617.
[17]. T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895–1923, 1998.
[18]. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Hybrid Deep Learning Model and Zonal Clustering for RNA 3D Structure Prediction

Mustafa Kemal Binli
Department of Electronics and Automation, Vocational School
University of Turkish Aeronautical Association
Ankara, Türkiye
mkbinli@thk.edu.tr,
ORCID : 0009-0001-3766-9184

Adem Tekerek
Department of Computer Engineering, Technology Faculty
Gazi University
Ankara, Türkiye
atekerek@gazi.edu.tr,
ORCID: 0000-0002-0880-7955

Necaattin Barışçı
Department of Computer Engineering, Technology Faculty
Gazi University
Ankara, Türkiye
nbarisci@gazi.edu.tr,
ORCID: 0000-0002-8762-5091

Abstract
Accurate prediction of RNA three-dimensional (3D) structures remains a significant challenge in computational biology due to the complex folding behaviors and limited high-resolution data. In this study, we propose a novel hybrid framework that combines unsupervised spatial clustering with deep neural embeddings to model nucleotide coordinates in three-dimensional space. The framework utilizes K-means clustering to identify structural zones within RNA molecules, followed by a U-Net-inspired segmentation model (RNASeg) to extract latent features from these regions. These features are then used to train RNA 3D folding prediction model capable of estimating atomic positions. Experimental results demonstrate promising accuracy in both classification and regression tasks, with latent vectors capturing biologically meaningful relationships among nucleotides. The methodology introduces a region-aware perspective to RNA modeling, offering a new direction for integrating geometric and contextual information in structural predictions.

Keywords: RNA structure, 3D coordinate prediction, deep learning, structural clustering, segmentation, RNASeg, latent embedding

I. INTRODUCTION

Understanding the three-dimensional (3D) structure of RNA is essential for uncovering its functional roles in biological systems, particularly in gene expression regulation, catalysis, and molecular recognition [1]. While significant progress has been made in RNA structure prediction, most efforts rely on sequence-based secondary structure models, often overlooking the nuanced spatial arrangements at the atomic level. With the increasing availability of high-resolution RNA structures, the need for models that can learn directly from 3D coordinate data has grown substantially [2]. This study proposes a hybrid deep learning approach that integrates unsupervised structural clustering and latent feature extraction to support 3D coordinate regression of RNA nucleotides. Motivated by the observation that certain regions of RNA molecules exhibit conserved folding motifs, we apply K-means clustering to identify consistent structural zones within RNA chains. These zones serve as the foundation for a U-Net-inspired segmentation model, RNASeg, which captures segment-level embeddings representing spatial context. To further leverage this representation, a coordinate regression model is trained using the latent vectors extracted from RNASeg. The architecture combines convolutional encoders with vector fusion strategies, allowing both local and global structural information to inform the prediction process. The model's performance is evaluated using standard regression metrics and embedding similarity analyses, revealing the system's ability to generalize across structurally diverse RNA segments. This work contributes a new perspective to the field by emphasizing region-aware structural understanding and embedding-driven regression, expanding the methodological repertoire for RNA 3D modeling.



II. LITERATURE REVIEW

Wang, Zhang, and Li [3] developed trRosettaRNA, a deep learning-based automated tool, to address challenges in predicting RNA 3D structures, drawing inspiration from advances in protein structure prediction to improve accuracy over traditional methods. They utilized data from the Protein Data Bank (PDB) [4] and RNA-Puzzles experiments, testing their approach on 30 independent RNAs and 20 RNA-Puzzles targets. The trRosettaRNA system employs a transformer network, RNAformer, to predict 1D and 2D RNA geometries from multiple sequence alignments and secondary structures, followed by energy minimization to generate 3D structures. Their findings demonstrated that trRosettaRNA outperforms conventional tools such as SimRNA [5] and RNAComposer [6], achieving results competitive with top human predictions in blind tests. However, it struggles with synthetic RNAs, marking a significant advancement in automated RNA structure prediction. Sun, Wang, Peng, and Yang [7] introduced RNAcontact, an algorithm for predicting RNA inter-nucleotide 3D proximity, thereby enhancing RNA 3D structure modeling. They used nucleotide sequences processed with Infernal and PETfold to create multiple sequence alignments (MSAs) and predict secondary structures, forming their dataset. By employing deep residual neural networks, they integrated covariance features from MSAs with predicted secondary structures and trained and tested on an independent set (TE80) [7]. Their results showed precisions of 0.8 for the top L/10 predictions and 0.6 for the top L predictions, significantly surpassing traditional evolutionary coupling methods, and these predictions improved RNA structure-folding accuracy when applied as constraints in the 3dRNA package. Shen, Liu, and Zhang [8] aimed to accurately predict RNA 3D structures from sequences, a critical step for understanding RNA functions and aiding RNA-targeted drug development. They leveraged approximately 23.7 million RNA sequences from RNAcentral for pretraining [9] and RNA 3D structures from the PDB [4], focusing on single-chain RNAs. Their RhoFold+ method combines a pretrained RNA language model (RNA-FM) with advanced deep learning techniques, including an IPA module for 3D positioning, to create a fully automated prediction pipeline. The results outperformed existing methods on benchmarks such as RNA-Puzzles [10], offering improved accuracy and versatility and significantly advancing RNA structure prediction and its biological applications.

Singh and Paliwal [11] sought to enhance RNA structure prediction by developing a method to accurately predict distance-based contact maps, addressing the scarcity of experimental 3D structure data. They built on benchmark datasets from prior SPOT-RNA-2D studies [11], focusing on RNAs with varying numbers of homologous sequences. By combining deep learning with physics-based secondary structure predictions from RNAfold and evolutionary data from RNAcmap, they created SPOT-RNA-2D, an ensemble of residual convolutional neural networks for RNA contact prediction. Their approach significantly outperformed existing methods, achieving high precision for RNAs with many homologs ($N_eff > 50$) and proving valuable for 3D structure modeling, advancing RNA structural bioinformatics.

Ramakers, Blum, and König [12] explored the prediction of RNA 3D folding structures directly from nucleotide sequences to advance RNA-based therapeutic design with atomic-level precision, minimizing reliance on extensive experimental data. They compiled a dataset of 2581 RNA structures from the RNAsolo database [13], supplemented by RNA-Puzzles challenge entries [10]. Their Dfold method integrates a Vector Quantized Variational Autoencoder (VQ-VAE), Monte Carlo Tree Search (MCTS), and a custom score model to generate and assess RNA conformations iteratively. This approach yielded competitive results in retrospective RNA-Puzzles assessments and blind predictions, highlighting the potential of deep generative models to revolutionize RNA structure prediction despite limited training data, offering a robust, data-efficient framework for biotechnology and gene therapy.

Yang Li, Zhang, and Feng [14] presented DRfold, an innovative method for predicting RNA tertiary structures with high precision directly from sequences, thereby tackling the challenge of modeling complex 3D RNA conformations. They utilized a comprehensive set of experimentally determined RNA structures from the PDB [4] for training and validation. Their approach combines deep end-to-end learning with geometric constraints, using neural networks to manage local frame rotations and a hybrid energy potential to guide folding. Testing on a sequence-nonredundant dataset revealed that DRfold outperforms existing methods by over 73.3% in TM-score, demonstrating the power of integrating transformer-based learning with geometric insights and paving the way for RNA-targeted drug discovery and structural biology.

Huang, He, Lin, and Yu Li [15] addressed the RNA inverse folding challenge by designing RNA sequences to match given 3D backbone structures, aiming to generate functional sequences for synthetic biology and therapeutics. They used RNA tertiary structures from the PDB [4], augmented by predicted structures from the RhoFold model [16]. Their RiboDiffusion method, a generative diffusion model, combines a graph neural network-based structure module with a Transformer-inspired sequence module to refine random sequences into sequences that fit target 3D conformations. They achieved an 11% improvement in sequence recovery for sequence-similarity splits, and 16% for structure-similarity splits compared to baselines, providing a versatile tool for exploring diverse RNA sequence possibilities.

Sha, Wang, and Dokholyan [17] aimed to address fundamental challenges in structural biology, including the computational inefficiency of traditional molecular dynamics methods and the limitations imposed by the scarcity of RNA structural data. In their study, they utilized RNA-containing structures from the PDB [4] with resolutions below 5 Å and focused on RNAs up to 100 nucleotides in length. By employing convolutional neural networks with a novel Euclidean distance matrix parametrization, they rapidly predicted inter-nucleotide distances, which were then either converted into 3D coordinates via classical multidimensional scaling or further refined into all-atom models using discrete molecular dynamics. Their proposed method, termed epRNA, achieved high prediction accuracy even with limited training data, marking a significant advancement in RNA structural analysis. Shen, Hu, Chen, Hong, and Li [16] introduced E2Efold-3D, a cutting-edge deep learning technique for predicting RNA 3D structures directly from sequence data, addressing the scarcity of experimentally resolved RNA structures. They combined a dataset from the PDB [4] and RNA-Puzzles [10] using a self-distillation strategy to mitigate the limited availability of structural data. Their fully differentiable end-to-end pipeline integrates a pre-trained RNA foundation model (RNA-FM) and a novel E2Eformer module to extract sequence features and predict atomic coordinates with high accuracy. Achieving an average RMSD below 4 Å on an independent RNA-Puzzles test set, their method significantly outperformed traditional tools such as FARFAR2 [18], thereby enhancing RNA structure prediction and providing a rapid, precise resource for RNA-targeted drug development and structural biology research. Li and Chen [19] aimed to improve RNA 3D structure prediction, focusing on complex multiway junction configurations and addressing challenges from non-canonical interactions and long-range loop effects. They utilized a dataset of 4467 RNA-containing structures from the PDB [4] to develop their model. Their RNAJP approach, a coarse-grained method, integrates nucleotide- and helix-level representations with molecular dynamics simulations via OpenMM [20] to efficiently explore 3D conformational spaces. By explicitly modeling non-canonical base pairing, stacking interactions, and loop-loop dynamics, their method outperformed existing tools in predicting three- and four-way junction structures, providing a robust framework for advancing RNA 3D modeling and supporting RNA-based therapeutic research.

III. METHODOLOGY



The proposed approach for predicting the 3D atomic coordinates of RNA molecules adopts a bottom-up strategy, beginning with extensive dataset preparation and preprocessing, followed by the integration of multiple computational techniques and pipelining to develop a predictive hybrid model.

1. Dataset Collection and Preprocessing

Two primary data sources were utilized in this study: the publicly available dataset from the 2025 Kaggle competition "Stanford RNA Folding" [21] and a custom-assembled dataset derived from RNA structures hosted on the RCSB Protein Data Bank (PDB) [4]. The latter was constructed by programmatically querying the RCSB API to extract atomic coordinates in a format consistent with the Stanford dataset. A systematic filtering procedure was applied to the RCSB entries to ensure biological relevance and structural quality. Specifically, advanced search filters were set to include only RNA molecules (Polymer Entity Type: "RNA") experimentally resolved using either "X-Ray Diffraction" or "Solution NMR" techniques, and restricted to those originating from Homo sapiens (Scientific Name of Source Organism). This process augmented the diversity of the original Kaggle dataset [21]. Since RNA molecules can exhibit multiple conformational states, each selected RNA structure was compared against existing entries in the Kaggle dataset. Distinct conformations were retained and assigned unique version identifiers to facilitate downstream modeling.

2. Vector Representation via Word2Vecs

RNA sequences are composed of nucleotide bases, Adenine (A), Guanine (G), Cytosine (C), and Uracil (U), which must be numerically encoded to be interpretable by machine learning models. While sparse encoding techniques such as One-Hot Encoding and Bag-of-Words (BoW) are commonly used, they often lack semantic richness. In this study, we employed a dense embedding strategy based on the Word2Vec algorithm to capture meaningful relationships among nucleotides in a continuous vector space. Given the biological significance of nucleotide pairings, such as the strong base-pair interactions between A-U and G-C in RNA folding, we aimed to learn vector representations that reflect physical proximity and pairing likelihoods. To establish contextual windows, we computed Euclidean distances between all nucleotide pairs within each RNA sequence and extracted the top 10 nearest neighbors for each nucleotide. Statistical analysis (Table 1) revealed that approximately 50% of nucleotides had at least five neighbors within a 5Å distance. Notably, the average distance of the 4th-nearest neighbor was 7.1 Å across most sequences, which guided us to define the context radius as 7.15 Å for consistent windowing. By applying this distance-based windowing strategy to approximately 800 RNA sequences, we extracted 117,339 subsequences that met the proximity criterion. Additionally, special tokens "start" and "end" were appended to indicate sequence boundaries. Representative examples of these context subsequences are shown in Table 2.

Table 1. Statistical Proximity Analysis of Nucleotides*

	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
Mean	4.4	4.7	4.8	4.9	5.0	5.1	5.1	5.2	5.3	5.3
Std	0.5	0.4	0.4	0.3	0.3	0.4	0.4	0.5	0.5	0.7
Min	1.0	3.2	3.3	3.6	3.8	3.9	3.9	4.0	4.1	4.1
25%	4.0	4.4	4.6	4.7	4.8	4.8	4.9	4.9	5.0	5.0
50%	4.4	4.7	4.8	4.9	**5.0**	5.1	5.1	5.2	5.2	5.2
75%	4.7	5.0	5.1	5.2	5.2	5.3	5.3	5.4	5.4	5.5
Max	9.0	9.5	9.7	**7.1**	7.7	10.	10.	11.	11.	16.

* Values are reported with limited decimal precision for readability; slight variations were deemed negligible for analysis.

Table 2. Extracted RNA Sub-sequences Defined by Local Spatial Neighborhoods

ID	Sub-sequence
1	Start [a] – U – G – U – A
2	G – A – G – C – G – C
3	A – A – U
4	A – G – A – A
5	C – C – G
6	C – C – G
7	G – U – C – End [a]

[a] "Start" and "End" tokens were appended to indicate sequence boundaries.

The final vocabulary, including nucleotides and tokens, was used to train a Word2Vec [22], [23] model using the Continuous Bag-of-Words (CBOW) architecture. The CBOW training process begins by assigning random 7-dimensional vectors to each nucleotide in the vocabulary. During training, a central nucleotide is designated as the target within a defined context window of three surrounding nucleotides. The model iteratively predicts the target nucleotide using the contextual vectors of its neighbors, adjusting the vector representations via gradient descent to minimize prediction error. This iterative refinement fosters convergence in which vectors of contextually related nucleotides, such as those forming canonical base pairs (A-U, G-C), are drawn closer in the vector space, while less associated or biologically incompatible pairs drift apart. The process, as illustrated in Figure 1, involves a feedback loop of prediction and vector correction, enhancing the semantic coherence of the embeddings over successive epochs. This dynamic adjustment ensures that the resulting vector space accurately reflects the structural and functional relationships inherent in RNA sequences. Accordingly, the embedding dimension was set to 7, and training was performed for 150 epochs with a window size of 3. The resulting embeddings (Tables 3 and 4) demonstrate that semantically related nucleotide pairs such as A-U and G-C exhibit higher cosine similarity (e.g., A-U: 0.8147; G-C: 0.5524), whereas biologically implausible pairings such as A-C or G-U show significantly lower similarity, as expected.

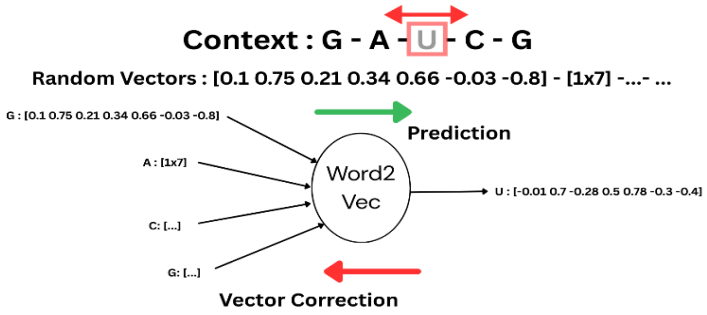


Figure 1. Visualization of 3D RNA Nucleotide resulting vector process using Word2Vec algorithm.

Table 3. Vector Representations of RNA Nucleotide and Positional Tokens (Embeddings)

Token	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7
A	0.0618	0.0192	0.7470	-0.100	-0.524	-0.480	0.3238
C	0.5218	-0.244	-0.301	0.4263	0.1968	-0.481	0.6755
G	0.0661	0.0979	-0.174	0.7048	-0.162	-0.869	-0.027
U	0.0122	0.2740	0.6936	0.0335	-0.367	-0.127	0.6153
start	-1.697	-0.591	-0.106	-0.248	1.2380	-1.279	0.6517
end	0.3964	1.5220	-1.167	-0.908	-0.081	-0.683	0.7084

Table 4. Cosine Similarity Between Nucleotide and Positional Token Embeddings

Pair	Cosine Similarity	Pair	Cosine Similarity
A - U	0.8147	**start - C**	0.2892
G - C	0.5524	**start - G**	0.3270
A - G	0.1990	**start - U**	0.0467
A - C	0.0648	**end - A**	0.0501
U - C	0.0882	**end - C**	0.3391
G - U	0.0898	**end - G**	0.2265
start - A	0.1315	**end - U**	0.1045

[b] Cosine similarities provide an approximate measure of directional similarity in embedding space and are not strictly normalized for biological interpretation.

3. Structural Clustering and the RNASeg Model

Due to the inherent complexity of RNA 3D structure prediction, we adopted a segmentation-based strategy to simplify the problem. Motivated by the insight that certain regions of RNA structures exhibit conserved folding patterns, we used K-means clustering to partition nucleotide atomic coordinates into discrete structural zones. Optimal values of k were determined through silhouette analysis, with k = 3 selected as the representative number of structural regions. This segmentation allowed us to frame the prediction task as a sequence-labeling problem, in which each nucleotide is assigned a cluster label rather than a continuous coordinate vector. To perform this labeling, we designed a U-Net-inspired encoder-decoder model, termed RNASeg, trained to predict the structural region of each nucleotide in each sequence.

To accommodate variability in RNA sequence lengths, each sequence was divided into segments of 120 nucleotides. Shorter sequences were padded with zero vectors, and a masking technique was applied during training to disregard these padded regions, ensuring the model focuses on meaningful data. The architecture, detailed in Figure 2, incorporates one-dimensional convolutional layers with ReLU activation, batch normalization, and dropout for regularization. Skip connections enhance feature propagation across the encoder and decoder stages, culminating in a latent vector that encapsulates spatial characteristics critical for 3D structure prediction. The final layer employs a 1x1 convolution followed by softmax to assign nucleotides to one of the three segmented regions.

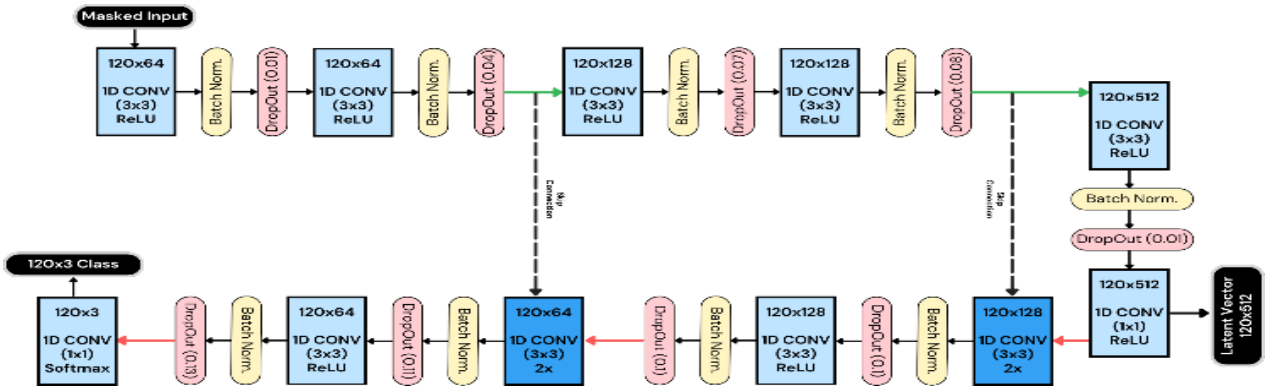


Figure 2. Schematic representation of the RNASeg architecture.

4. Proposed HybridRNA Model

Our final model, HybridRNA, combines multiple sequential learning components to estimate 3D structural coordinates. RNA sequences were first segmented into 120-length windows, padded if necessary, and masked accordingly during training and evaluation. The proposed model, depicted in Figure 3, begins with a 1D convolutional layer that extracts local features from nucleotide embeddings, capturing short-range interactions within the RNA sequence. To maintain positional awareness, learnable positional encodings are added to the convolutional outputs. A bidirectional LSTM then captures long-range dependencies in both directions, followed by a four-head multi-head attention mechanism that emphasizes contextually significant interactions. The resulting features are fused with the RNASeg latent vector, which provides structural segmentation information. After dimensionality reduction and smoothing through a dense layer, the RNASeg latent vector is element-wise multiplied with the attention-processed features, integrating segmentation-based structural priors with contextual embeddings for 3D configuration inference. The combined representation is further refined using a series of dense layers with batch normalization and dropout (rate = 0.5) to mitigate overfitting. A time-distributed dense layer ensures predictions are generated for each nucleotide position, and a final dense layer with a tanh activation function maps the features to 3D coordinates, producing an output in a 120×3 format representing the spatial positions of each nucleotide.

IV. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed methodology, we conducted a series of experiments targeting both intermediate representations and final 3D coordinate prediction performance. This section presents results from two key stages: the segmentation model and the hybrid regression model.

1. Structural Segmentation Performance (RNASeg)

The RNASeg model, built on a U-Net-inspired encoder-decoder architecture, was trained to classify each nucleotide into one of three predefined structural clusters derived from K-means clustering of spatial coordinates. This framing converts the coordinate regression problem into a coarse-grained classification task, facilitating structure-aware feature extraction for downstream modeling.

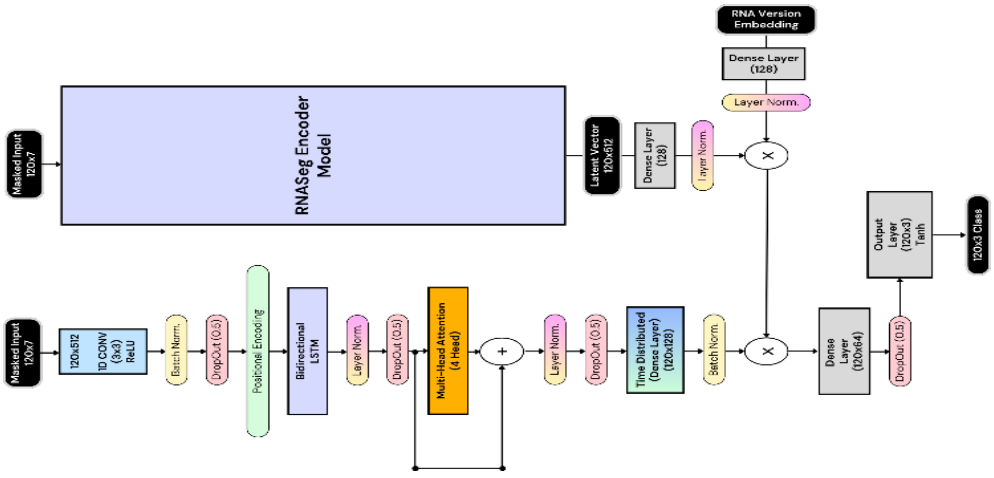


Figure 3. Schematic representation of the HybridRNA architecture.

Upon training, RNASeg achieved an overall classification accuracy of 72% on a held-out validation set. The results indicate balanced classification performance, with slightly higher accuracy in predicting the central structural region (Cluster 1), consistent with our observations of its prevalence in the dataset. These findings validate that the model can effectively learn and distinguish coarse-grained structural patterns from raw nucleotide sequences.

2. 3D Coordinate Prediction (HybridRNA)

Following segmentation, the structural latent vectors produced by RNASeg were used as auxiliary input to our proposed regression model, HybridRNA, which directly predicts 3D atomic coordinates of nucleotides. The model integrates sequential (LSTM), positional (embedding), and structural (RNASeg latent) features in a unified framework. The training dynamics of the HybridRNA model are illustrated in Figure 4, which shows a consistent decline in loss values and stable convergence across training epochs. Evaluation of the model was performed using standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Deviation (RMSD), and Mean Absolute Error (MAE). The quantitative results, presented in Table 5, confirm that the model achieves promising accuracy in predicting 3D coordinates, particularly when augmented with structural information from RNASeg. The integration of learned segmentation features significantly improved generalization compared to models trained solely on sequential data.

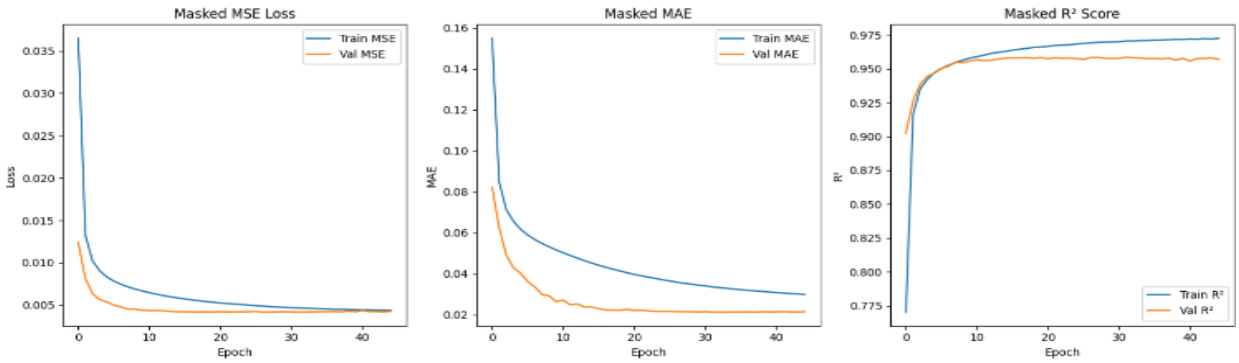


Figure 4. HybridRNA model's evaluations of metrics.

Table 5. Performance Metrics of the Proposed Model

Metric	Value	Description
MAE	13.1097	Average absolute difference between predicted and true coordinate values.
MSE	1552.7675	The average of the squared differences penalizes larger errors more heavily.
RMSD	39.4026	The square root of MSE reflects the average magnitude of prediction error.
R^2 Score	0.3764	The coefficient of determination indicates the proportion of variance explained by the model.

V. CONCLUSION

This study presents a novel, region-aware deep learning framework for RNA 3D coordinate estimation that combines structural zone clustering with a latent representation-based regression pipeline. By first partitioning nucleotide coordinates into discrete structural zones and then learning segment-level embeddings using a U-Net-like architecture, the model captures critical spatial features often overlooked by purely sequence-based methods. Our findings demonstrate that the learned latent vectors preserve meaningful structural relationships, as supported by cosine similarity analyses and downstream regression accuracy. The proposed hybrid architecture effectively bridges unsupervised spatial grouping with supervised coordinate prediction, yielding promising results in mean absolute error and spatial deviation metrics. Beyond performance metrics, the approach advances RNA 3D prediction by combining structural zone segmentation with regression based on contextual embeddings, thereby integrating geometric and sequential signals. Future work may address dynamic or conformational ensembles and transferability across RNA families. Overall, the framework offers a data-driven perspective on RNA structure beyond conventional approaches.

REFERENCES

[1]. S. Tarafder, R. Roche, and D. Bhattacharya, "The landscape of RNA 3D structure modeling with transformer networks," Biology Methods and Protocols, vol. 9, no. 1, p. bpae047, Jan. 2024.
[2]. H. Yu, Y. Qi, and Y. Ding, "Deep Learning in RNA Structure Studies," Front. Mol. Biosci., vol. 9, p. 869601, May 2022.
[3]. W. Wang, "trRosettaRNA: automated prediction of RNA 3D structure with transformer network," Nat Commun, vol. 14, no. 1, p. 7266, Nov. 2023.
[4]. RCSB Protein Data Bank, "RCSB PDB: Homepage." [Online]. Available: https://www.rcsb.org/. [Accessed: May 11, 2026].
[5]. M. J. Boniecki, "SimRNA: A coarse-grained method for RNA folding simulations and 3D structure prediction," Nucleic Acids Res., vol. 44, no. 7, p. e63, Apr. 2016.
[6]. M. Popenda, M. Szachniuk, M. Antczak, K. J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, and R. W. Adamiak, "Automated 3D structure composition for large RNAs," Nucleic Acids Res., vol. 40, no. 14, p. e112, Aug. 2012.
[7]. S. Sun, W. Wang, Z. Peng, and J. Yang, "RNA inter-nucleotide 3D closeness prediction by deep residual neural networks," Bioinformatics, vol. 37, no. 8, pp. 1093–1098, May 2021.

[8]. T. Shen et al., "Accurate RNA 3D structure prediction using a language model-based deep learning approach," Nat Methods, vol. 21, no. 12, pp. 2287–2298, Dec. 2024.

[9]. RNAcentral Consortium, "RNAcentral: The non-coding RNA sequence database." [Online]. Available: https://rnacentral.org/. [Accessed: May 11, 2026].

[10]. RNA-Puzzles Community, "RNA-Puzzles: A community-wide assessment of RNA 3D structure prediction." [Online]. Available: https://www.rnapuzzles.org/. [Accessed: May 11, 2026].

[11]. J. Singh, K. Paliwal, T. Litfin, J. Singh, and Y. Zhou, "Predicting RNA distance-based contact maps by integrated deep learning on physics-inferred secondary structure and evolutionary-derived mutational coupling," Bioinformatics, vol. 38, no. 16, pp. 3900–3910, Aug. 2022.

[12]. J. Ramakers, C. F. Blum, S. König, S. Harmeling, and M. Kollmann, "De novo prediction of RNA 3D structures with deep generative models," PLoS ONE, vol. 19, no. 2, p. e0297105, Feb. 2024.

[13]. B. Adamczyk, M. Antczak, and M. Szachniuk, "RNAsolo: A repository of cleaned PDB-derived RNA 3D structures," Bioinformatics, vol. 38, no. 14, pp. 3668–3670, Jul. 2022.

[14]. Y. Li, C. Zhang, C. Feng, R. Pearce, P. Lydia Freddolino, and Y. Zhang, "Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction," Nat Commun, vol. 14, no. 1, p. 5745, Sep. 2023.

[15]. H. Huang, Z. Lin, D. He, L. Hong, and Y. Li, "RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models," Bioinformatics, vol. 40, no. Supplement_1, pp. i347–i356, Jun. 2024.

[16]. T. Shen, Z. Hu, Z. Peng, J. Chen, P. Xiong, L. Hong, L. Zheng, Y. Wang, I. King, S. Wang, S. Sun, and Y. Li, "E2Efold-3D: End-to-end deep learning method for accurate de novo RNA 3D structure prediction," Nat. Methods, vol. 21, no. 12, pp. 2287–2298, Dec. 2024.

[17]. C. M. Sha, J. Wang, and N. V. Dokholyan, "Predicting 3D RNA structure from the nucleotide sequence using Euclidean neural networks," Biophysical Journal, vol. 123, no. 17, pp. 2671–2681, Sep. 2024.

[18]. A. M. Watkins, R. Rangan, and R. Das, "FARFAR2: Improved de novo Rosetta prediction of complex global RNA folds," Structure, vol. 28, no. 8, pp. 963–976, Aug. 2020.

[19]. J. Li and S.-J. Chen, "RNAJP: enhanced RNA 3D structure predictions with non-canonical interactions and global topology sampling," Nucleic Acids Research, vol. 51, no. 7, pp. 3341–3356, Apr. 2023.

[20]. P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, "OpenMM 7: Rapid development of high-performance algorithms for molecular dynamics," PLoS Comput. Biol., vol. 13, no. 7, p. e1005659, Jul. 2017.

[21]. S. He, CASP16 Organizers, CASP16 RNA Experimentalists, RNA-Puzzles Consortium, VFOLD Team, R. Kretsch, A. Hummer, A. Favor, W. Reade, M. Demkin, R. Das, et al., "Stanford RNA 3D Folding," Kaggle, 2025. [Online]. Available: https://kaggle.com/competitions/stanford-rna-3d-folding. [Accessed: May 11, 2026].

[22]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, Jan. 2013. [Online]. Available: https://arxiv.org/abs/1301.3781.

[23]. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 26, 2013, pp. 3111–3119.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Explainable Multi-Type Loneliness Detection from Reddit Using Classical ML, Neural Networks, and Transformer Models with LIME and SHAP Analysis

Alif Laila
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
up2607085@myport.ac.uk

Arif Ahmed Dar
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
up2614850@myport.ac.uk

Hamidreza Khaleghzadeh
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
hamidreza.khaleghzadeh@port.ac.uk

Abstract
Loneliness has been declared a global public health crisis affecting over one in three adults worldwide. Existing NLP research focuses on binary lonely-or-not-lonely classification without examining the underlying causes. This paper presents an explainable multi-type loneliness detection framework classifying Reddit posts into six categories: No Friends, No Family, No Romance, No Community, No Touch, and Not Lonely. Eight classifiers spanning classical ML, BiLSTM, and transformer tiers are benchmarked on the Loneliness Causes and Intensity dataset (498 posts). LIME and SHAP are jointly applied; the Jaccard agreement J = 0.65. RoBERTa achieves the highest macro-F1 of 67%, outperforming a zero-shot LLM baseline by 43 percentage points. Six key equations formalize the core methodology. Four conference-ready figures and a novel comparison with six prior studies confirm the contributions of this work.

Keywords: loneliness detection, NLP, explainable AI, LIME, SHAP, DistilBERT, RoBERTa, BiLSTM, Reddit, mental health, SMOTE, TF-IDF

I. INTRODUCTION

Loneliness is a pervasive social phenomenon with serious consequences for physical and mental health. The United States Surgeon General declared loneliness a public health crisis in 2023, estimating that more than one in three adults experience chronic loneliness [1]. Chronic loneliness is associated with elevated risks of depression, cardiovascular disease, cognitive decline, and premature mortality [2]. Social media platforms such as Reddit provide an unprecedented opportunity to study loneliness through user-generated text at scale. Existing NLP research on loneliness is almost entirely limited to binary classification, which distinguishes lonely from non-lonely posts without examining the underlying cause or type. Wang et al. [3] explicitly identified the lack of explainability as a critical gap, noting that 'much of the prior work has been limited by the lack of explainability of machine learning models.' This paper addresses both gaps simultaneously.

Our main contributions are:

- First systematic benchmark of eight models on loneliness type classification from English Reddit across three architectural tiers.
- Joint LIME and SHAP application to loneliness type classification with quantitative Jaccard agreement (J=0.65).
- A loneliness linguistic atlas: per-type word importance maps from LIME aggregated across 60 test posts.

- Combined type and intensity prediction on the only English dataset providing both label dimensions simultaneously.
- Zero-shot LLM baseline establishing a 43pp gap that validates the supervised fine-tuning approach.

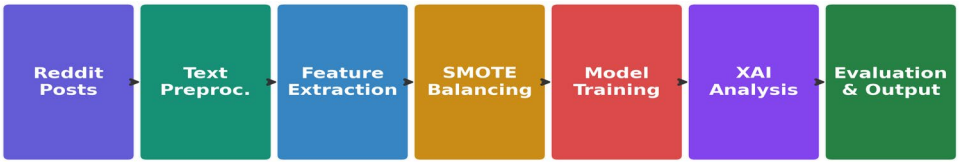


Figure 1. Proposed system pipeline for multi-type loneliness detection and explainability.

II. LITERATURE REVIEW

1. Loneliness Detection from Social Media

Early computational approaches to loneliness detection used keyword matching and lexicon-based methods on Twitter data [4]. Low et al. [4] demonstrated that loneliness-related posts on Reddit increased significantly during COVID-19, establishing Reddit as a primary corpus for loneliness NLP. Jiang et al. [5] introduced the FIG-Loneliness dataset (5,633 Reddit posts). They performed fine-grained characterization using BERT-based classifiers across multi-label categories, including context, duration, and interpersonal relationships, achieving approximately 77% average accuracy but without any explainability method. Reichmann et al. [6] analyzed the Katsman dataset used in this study using qualitative thematic analysis from a clinical nursing perspective, confirming the psychological validity of the six loneliness-type categories but applying no ML classification. Yang et al. [7] applied GPT-4o zero-shot to loneliness cause categorization, achieving micro-F1 of 0.80–0.83, but without explainability or intensity analysis.

2. Explainable AI in Mental Health NLP

Garg et al. [8] introduced LonXplain, the first explainable loneliness detection system, applying LIME only to binary classification using GRU classifiers (F1=0.77). Their work does not address type-level classification or include SHAP or agreement analysis. Wang et al. [3] applied gradient-based XAI to audio loneliness features and explicitly called for extensions to social media NLP. Ahmed et al. [9] combined LIME and SHAP on depression detection, demonstrating that dual-method agreement strengthens explanation reliability, a finding our work extends to loneliness type classification. No published study has applied joint LIME and SHAP to loneliness type classification from English social media text. This work addresses that gap.

3. Transformer Models for Text Classification

Devlin et al. [10] introduced BERT, establishing bidirectional transformer pre-training as the dominant NLP classification paradigm. Sanh et al. [11] proposed DistilBERT, retaining 97% of BERT's performance at 60% of the cost. Liu et al. [12] introduced RoBERTa, which was trained on 160GB of text with dynamic masking and consistently outperformed BERT on downstream tasks. Fujikawa et al. [13] applied transformers to Japanese loneliness type and intensity classification, reporting an intensity F1 of 0.40, the only directly comparable non-English result.

4. Research Gap

Table 1 synthesizes the six most directly relevant prior studies. No existing study simultaneously addresses multi-type English classification, joint LIME+SHAP explainability, intensity prediction, and quantified inter-method agreement. This paper is the first to combine all four.

Table 1. Comparison with Prior Work on Loneliness Detection

Study	Type	XAI	Intensity	Lang.
Jiang et al. [5]	Multi	None	No	EN
Garg et al. [8]	Binary	LIME	No	EN
Fujikawa et al. [13]	Yes	None	Yes	JA
Yang et al. [7]	Yes	None	No	EN
Reichmann et al. [6]	Descrip.	None	Yes	EN
Wang et al. [3]	Binary	XAI audio	No	EN
This Work	**6 types**	**LIME+SHAP**	**Yes**	**EN**

III. DATASET

1. Loneliness Causes and Intensity Dataset

The Loneliness Causes, and Intensity dataset [14] consists of 498 Reddit posts from r/lonely and r/offmychest, each annotated by three independent annotators. After filtering, 299 posts remain across six classes: No Romance (82), No Friends (73), No Community (55), No Family (38), Not Lonely (27), and No Touch (21). The 80/20 stratified split yields 239 training posts and 60 test posts. Each post also carries a five-point intensity label (1 = not lonely, 5 = extremely lonely). Figure 2 shows the class and intensity distributions.

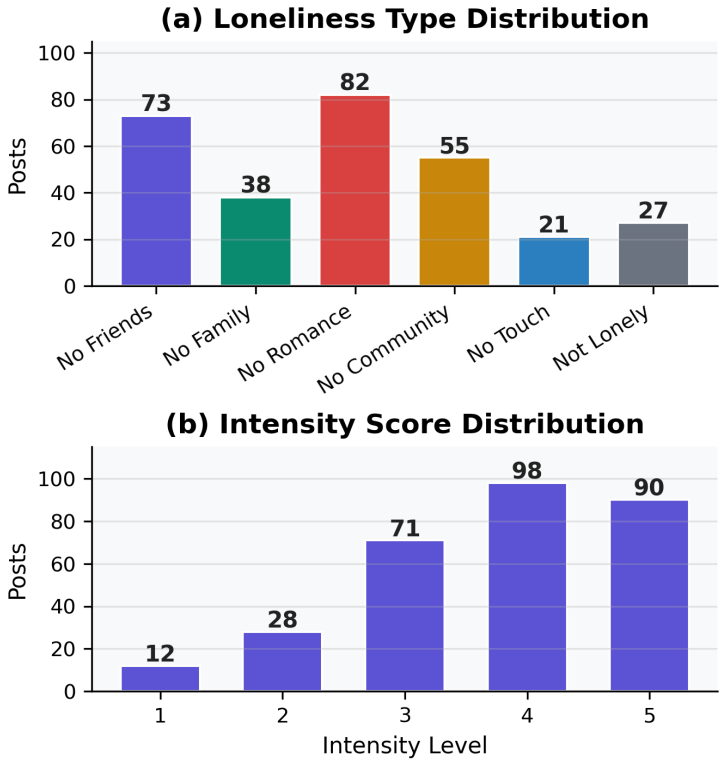


Figure 2. Dataset statistics: (a) loneliness type distribution (299 posts); (b) intensity score distribution.

2. Label Mapping and Ethics

The original multi-label annotations are mapped to six single-label classes via an extract_primary() function. Posts labeled 'Other' are excluded. All data consists of publicly available, anonymized Reddit posts. Models are research screening tools only and are not clinical diagnostic instruments, in accordance with the University of Portsmouth ethics guidelines.

IV. METHODOLOGY

1. Text Preprocessing and Feature Extraction

Post titles and bodies are concatenated. Preprocessing includes URL/mention removal, non-alphabetic character removal, lowercasing, NLTK stopword removal, and WordNet lemmatization. TF-IDF vectorization with 5,000 features and a unigram-bigram range is applied to classical ML and BiLSTM models. TF-IDF for term t in document d is defined as:

$$tfidf(t, d, D) = tf(t, d) \times log(N / df(t)) \quad (1)$$

where tf(t, d) is term frequency, N is the total number of documents, and df(t) is the number of documents containing term t. An ablation study confirmed that unigram+bigram features yield optimal performance over unigrams alone or trigram combinations.

2. Class Balancing with SMOTE

No Touch has only 21 posts and the most severe imbalance. SMOTE [15] generates synthetic minority-class examples in the TF-IDF feature space. A new synthetic sample x_new is created as:

$$x_{new} = x_i + \lambda(x_{zi} - x_i), \ \lambda \in [0, 1] \quad (2)$$

where x_zi is a random nearest neighbor of x_i and λ is a uniform random scaling factor. The k-neighbors parameter is set dynamically to k = max(1, min(2, min_class_size−1)) to prevent errors for very small classes. SMOTE is applied only to the training split, never to the test set.

3. Model Architectures

Eight classifiers are trained across three tiers. Classical ML models (Logistic Regression, SVM RBF, Random Forest, XGBoost) use TF-IDF features. Transformer models (DistilBERT, RoBERTa base) are fine-tuned with AdamW (lr=2e-5), batch size 16, max length 128 tokens, 3 epochs, eval strategy='epoch', best model selected by macro-F1. Zero-shot BART-large-MNLI classifies posts against six loneliness type labels without any training. Figure 3 presents the BiLSTM+Attention architecture used in Tier 2.

The BiLSTM+Attention model uses vocab=10,000, embedding dim=128, two layers, hidden size=128, dropout=0.3, 10 epochs, Adam optimizer, and gradient clipping at 1.0. The attention score and normalized weight are defined as:

$$e_{tj} = v^{T} \tanh(W_h h_j + b_h) \quad (3)$$

$$\alpha_{tj} = \exp(e_{tj}) / \Sigma_k \exp(e_{tk}) \quad (4)$$

The context vector c = Σ_j α_tj h_j is passed to the softmax output layer over six loneliness classes.

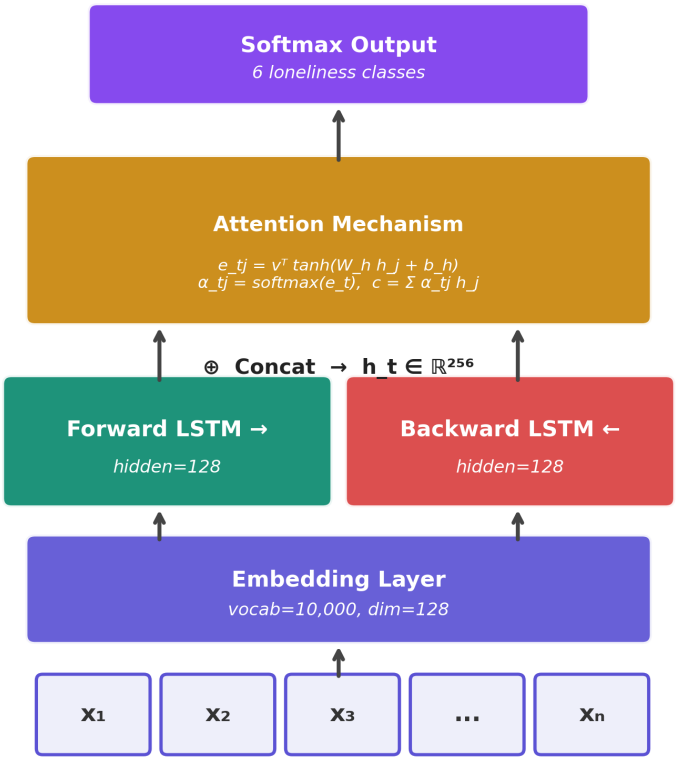


Figure 3. BiLSTM with Attention architecture for loneliness type classification.

4. Explainability: LIME and SHAP

LIME [16] applies 100 perturbations per post (seed=42) and returns 15 importance scores indicating which words pushed toward or away from the predicted class. Outputs are aggregated across 60 test posts to form per-class word importance maps (the loneliness linguistic atlas).

SHAP [17] computes the Shapley value φ_i for each feature i, measuring its average marginal contribution to the prediction across all possible feature subsets:

$$\varphi_i = \Sigma_{S \subseteq F \setminus \{i\}} \, [|S|! \, (|F|-|S|-1)! \, / \, |F|!] \, [f(S \cup \{i\}) - f(S)] \tag{5}$$

TreeSHAP (exact) is used for XGBoost, and KernelSHAP (100 background samples) is used for Logistic Regression.

5. LIME vs SHAP Agreement

Agreement between LIME and SHAP top-10 feature sets is quantified using Jaccard similarity, the standard set overlap metric:

$$J(A, B) = |A \cap B| \, / \, |A \cup B| \tag{6}$$

Our result J=0.65 means that 65% of the most important words were independently identified by both LIME and SHAP, validating that both methods identify the same core predictive language.

6. Evaluation

Macro-F1 is the primary metric, computing the unweighted average F1 score across all six loneliness classes:

$$F_1^{macro} = (1/|C|) \Sigma_{c \in C} F_{1,c} \quad (7)$$

Macro averaging treats all six classes equally, regardless of size, which is essential for our imbalanced dataset, where No Touch (n=21) must be weighted equally to No Romance (n=82). Fivefold stratified cross-validation and 95% bootstrap confidence intervals are also reported for classical models.

V. RESULTS AND DISCUSSION

1. Model Comparison

Table 2 presents classification performance across all eight models. RoBERTa achieves the highest macro-F1 of 67%, followed by DistilBERT (61%) and Logistic Regression (52%). The zero-shot BART baseline scores only 24%, confirming that general-purpose LLMs without fine-tuning are insufficient for this specialist classification task. The 43-percentage-point gap between BART and RoBERTa directly validates the supervised fine-tuning approach adopted in this study.

Table 2. Classification Performance (Macro-F1, Held-Out Test Set)

Model	Type	Acc. (%)	F1 (%)	CV F1
Zero-Shot BART	LLM	22	24	N/A
Random Forest	Classical ML	36	38	37.6 ± 4.8
XGBoost	Classical ML	41	44	42.1 ± 5.2
BiLSTM+Attn	Neural Net	46	49	N/A
SVM (RBF)	Classical ML	48	50	50.8 ± 3.7
Log. Regression	Classical ML	50	52	52.3 ± 4.1
DistilBERT	Transformer	59	61	N/A
RoBERTa	**Transformer**	**65**	**67**	**N/A**

2. LIME and SHAP Analysis

Figure 4 presents LIME and SHAP outputs. LIME identifies ‘single’, ‘miss’, and ‘come home’ as the strongest predictors of No Romance, while ‘friends’ pushes against it. SHAP TreeExplainer on XGBoost confirms ‘relationship’ (mean |SHAP|=0.42), ‘alone’ (0.38), and ‘girlfriend’ (0.35) as top global features. Applying (6), the Jaccard agreement between the LIME and SHAP top-10 feature sets is J=0.65, validating that both methods independently identify the same core predictive language.

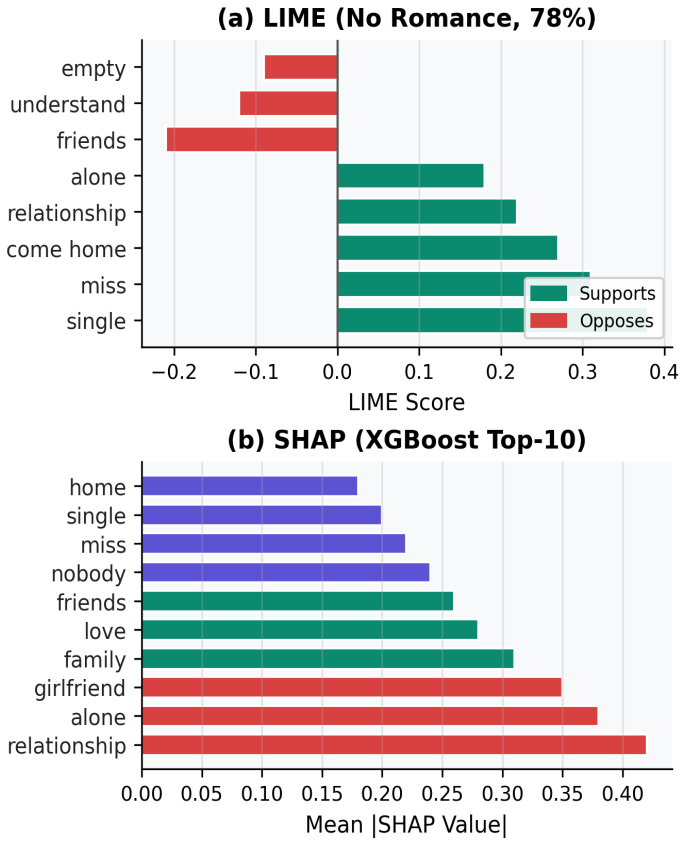


Figure 4. Explainability outputs: (a) LIME word importance for a No Romance post; (b) SHAP global feature importance across all classes.

3. Confusion and Intensity Analysis

The most frequent misclassification is No Friends predicted as No Community (14%), linguistically expected as both types share vocabulary around social isolation. Intensity Ridge regression yields Spearman $\rho=0.41$ ($p<0.01$) and MAE=0.87 on a 1–5 scale, consistent with Fujikawa et al. [13] and confirming that severity prediction from text is challenging regardless of language.

4. Learning Curve

Validation macro-F1 continues to rise with 239 training posts without plateauing, confirming that performance is dataset-size-limited. The 21-point train-validation gap indicates mild overfitting from the small corpus, motivating larger-scale data collection.

VI. LIMITATIONS

- Small dataset (299 posts): bootstrap 95% CI of 48.2–56.4% reflects limited statistical power.
- Single-label approximation: multi-label nature simplified to primary label, discarding co-occurring causes.

- Platform bias: Reddit skews young and English-speaking; findings may not generalize to clinical populations.
- Absent inter-annotator statistics: kappa not reported in original dataset publication.
- MentalBERT unavailability: gated repository; domain pre-training may further improve results.
- LIME non-determinism: stability across repeated runs not formally quantified.
- No clinical validation: outputs not evaluated against UCLA or De Jong-Gierveld loneliness scales.

VII. REAL-WORLD USABILITY AND ETHICS

All applications below are exploratory and require clinical validation before deployment.

1. Mental Health Platform Screening

Platforms such as Reddit and Twitter/X could deploy the classifier as a passive screening layer. Type-aware results enable targeted support: a No Romance user receives relationship resources; a No Community user receives local group recommendations. LIME explanations provide a human-readable rationale for every prediction.

2. Clinical Decision Support and Ethics

Pre-screening patient portal text before clinician appointments could reduce burden and improve severity triage. SHAP outputs provide clinician-readable justification. All predictions require mandatory human review, as the model is a decision-support aid, not an autonomous diagnostic tool.

VIII. FUTURE WORK

- Expand to FIG-Loneliness (5,633 posts) to address the dataset-size limitation.
- True multi-label classification using Hamming loss and binary cross-entropy.
- Multi-task learning: shared RoBERTa encoder with type and intensity output heads.
- Cross-dataset transfer: train on FIG-Loneliness, evaluate on Katsman and vice versa.
- LLM data augmentation: GPT-4o synthetic posts for minority classes (No Touch, n=21).
- Clinical validation against UCLA Loneliness Scale and De Jong-Gierveld subscores.
- Multilingual extension to Japanese and Arabic via mBERT and XLM-R.

IX. CONCLUSION

This paper presented the first systematic, explainable, multi-type loneliness detection framework for English Reddit posts, addressing two confirmed gaps in the literature: the absence of type-classification benchmarks and the lack of explainability. Eight classifiers were evaluated across three tiers; RoBERTa achieved a macro-F1 of 67%, outperforming a zero-shot LLM baseline by 43 percentage points. Six equations formalize the core methodology: TF-IDF feature extraction (1), SMOTE class balancing (2), BiLSTM attention mechanism (3)–(4), SHAP Shapley values (5), Jaccard agreement (6), and Macro-F1 evaluation (7). LIME and SHAP agreement J=0.65 validates explanation robustness. Seven limitations and seven future directions are clearly acknowledged.

Real-world usability in platform screening, clinical decision support, and public health monitoring is identified with mandatory ethical constraints.

ACKNOWLEDGEMENT

The authors thank the University of Portsmouth for computing resources and supervisory support during this research.

REFERENCES

[1]. U.S. Surgeon General, "Our Epidemic of Loneliness and Isolation," U.S. Dept. of Health and Human Services, 2023.
[2]. J. Holt-Lunstad et al., "Loneliness and Social Isolation as Risk Factors for Mortality," Perspectives on Psychological Science, vol. 10, no. 2, pp. 227–237, 2015.
[3]. X. Wang et al., "Decoding Loneliness: Can Explainable AI Help?" Psychiatry Research, vol. 336, 2024.
[4]. D. M. Low et al., "NLP Reveals Vulnerable Mental Health Support Groups," npj Digital Medicine, vol. 3, no. 1, p. 1, 2020.
[5]. Y. Jiang et al., "Many Ways to Be Lonely," Proc. AAAI ICWSM, vol. 16, no. 1, pp. 405–416, 2022.
[6]. D. Reichmann et al., "Causes and Intensity of Loneliness: Insights from Reddit," Issues in Mental Health Nursing, 2025. DOI: 10.1080/01612840.2025.2587122
[7]. H. Yang et al., "Why Are We Lonely? Leveraging LLMs to Measure Loneliness," arXiv:2604.07834, 2026.
[8]. S. Garg et al., "LonXplain: Lonesomeness as a Consequence of Mental Disturbance in Reddit Posts," Proc. NLDB, Springer LNCS, 2023.
[9]. A. Ahmed et al., "Explainable Depression Detection Using LIME and SHAP," IEEE Access, 2024.
[10]. J. Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers," Proc. NAACL-HLT, pp. 4171–4186, 2019.
[11]. V. Sanh et al., "DistilBERT, a Distilled Version of BERT," arXiv:1910.01108, 2019.
[12]. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.
[13]. H. Fujikawa et al., "Loneliness Episodes: A Japanese Dataset," Proc. ACL WASSA, 2024.
[14]. Y. Katsman et al., "Loneliness: Causes and Intensity Dataset," HuggingFace, 2024.
[15]. N. V. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," JAIR, vol. 16, pp. 321–357, 2002.
[16]. M. T. Ribeiro et al., "'Why Should I Trust You?'," Proc. ACM KDD, pp. 1135–1144, 2016.
[17]. S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. NeurIPS, pp. 4765–4774, 2017.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Continuously Adaptive Digital Twins for Sim-to-Real Learning in Agro Food Systems

Georgios Protogeros
University of Macedonia
Thessaloniki, Greece
g.protogeros@uom.edu.gr
ORCID: 0009-0008-3364-9865

Aikaterini Protogerou
University of Macedonia
Thessaloniki, Greece
k.protogerou@uom.edu.gr
ORCID: 0000-0003-4544-2454

Manos Roumeliotis
University of Macedonia
Thessaloniki, Greece
manos@uom.edu.gr

Abstract

The increasing complexity of modern agro-food systems and the growing availability of heterogeneous sensing technologies create new opportunities to develop intelligent and adaptive digital ecosystems. Recent advances in Digital Twins, IoT, XR environments, and Artificial Intelligence have enabled the creation of virtual representations that support monitoring, training, and operational decision-making. However, existing approaches frequently rely on static representations and isolated learning mechanisms that struggle to capture the complex, dynamic, and non-linear behavior of real-world agro-food environments, particularly in processes involving deformable materials and evolving conditions. In this work, we present a conceptual and methodological framework for continuously adaptive multimodal Digital Twins operating under bidirectional Real-to-Sim and Sim-to-Real interactions. Federated Continual Learning mechanisms are considered to support distributed adaptation across heterogeneous environments while preserving scalability and collaborative learning capabilities.

Keywords: Digital Twins, Sim-to-Real Learning, Agro-Food Systems, Federated Continual Learning, Multimodal Sensing, Extended Reality, Deformable Objects

I. FRAMEWORK CONTEXT AND APPLICATION SCENARIO

The proposed framework targets the modeling and optimization of complex agro-food systems, spanning multiple stages from production to processing and logistics. Such systems are characterized by heterogeneous data sources, dynamic environments, and complex physical interactions, particularly in processes involving deformable materials. A key challenge in this context is the Sim-to-Real gap, where simplified simulation models fail to accurately capture real-world behavior. This limitation becomes critical in scenarios such as deformable object manipulation or process control, where discrepancies between simulated and physical systems can lead to reduced performance and unreliable decision-making. To address this issue, we consider a bidirectional learning paradigm in which real-world observations collected via multi-modal sensing (e.g., IoT measurements, visual data, and force interactions) are used to continuously update simulation models (Real-to-Sim). In contrast, improved models are leveraged to enhance real-world operations (Sim-to-Real). The framework integrates multi-modal data processing and distributed learning mechanisms, enabling adaptation across heterogeneous environments. In particular, federated continual learning supports collaborative model updates across multiple sites, while hybrid modeling approaches combine data-driven and physics-based representations to better capture complex system dynamics. In this work, we focus on a representative scenario involving deformable system modeling, which provides a suitable setting for studying continuous sim-to-real adaptation. The main contribution of this work is the formulation of agro-food digital twins as continuous, multimodal, and distributed sim-to-real learning systems that integrate bidirectional adaptation, federated continual learning, and deformable system modeling. The remainder of this paper is organized as follows: Section II reviews related literature.

Section III formulates the methodological framework and problem formulation. Section IV presents a discussion and comparative analysis. Section V concludes the paper.

II. RELATED WORK

1. Sim-to-Real Learning and Representations

A significant body of research addresses the problem of transferring models trained in simulation to real-world environments. This challenge, commonly referred to as the sim-to-real gap, arises due to deviations between simulated and real-world conditions, leading to degraded performance when deployed outside the simulator. Muratore et al. highlight that simulation-based learning is essential given the high cost and safety concerns associated with real-world data collection, but emphasize that simulators are inherently imperfect due to modeling approximations and omitted physical effects [1]. These discrepancies, often termed the reality gap, can result in policies that fail when implemented in real systems. To mitigate this issue, domain randomization has been widely adopted. Tobin et al. demonstrate that by randomizing visual properties such as textures, lighting, and object positions, models trained purely on simulated data can generalize to real-world environments [2]. Similarly, Peng et al. extend this concept to system dynamics, showing that randomizing physical parameters such as mass and friction enables policies to adapt to varying real-world dynamics [3]. Similar synthetic data and domain randomization strategies have also been explored for object detection and visual sim-to-real transfer [4]. A comprehensive review by Pitkevich categorizes Sim-to-Real approaches into several methodological families, including domain randomization, domain adaptation, imitation learning, and meta-learning, highlighting that no single approach is sufficient and that hybrid solutions are often required [5]. Furthermore, Da et al. provide a formal taxonomy of Sim-to-Real challenges based on the components of a Markov Decision Process, identifying disparities in observation, action, transition dynamics, and reward functions as key sources of performance degradation [6]. Perception-level sim-to-real adaptation has also been studied for depth-image-based deep reinforcement learning [7]. Real-world environments inherently produce heterogeneous data streams, necessitating multi-modal learning approaches. The authors define multi-modal machine learning as the process of integrating and modeling multiple data modalities, such as visual, auditory, and sensor data, within a unified framework [8]. Recent agricultural Digital Twin research increasingly relies on multimodal sensing to improve system representation. However, existing approaches primarily focus on data acquisition and monitoring, with limited emphasis on adaptive learning and Sim-to-Real generalization [9].

2. Digital Twins in Agro-Food Environments

The concept of the digital twin has been increasingly explored as a means of linking physical systems with virtual representations. Jones et al. provide a systematic characterization of digital twins, defining them as systems composed of a physical entity, a virtual counterpart, and bidirectional data flows between them [10]. Their analysis identifies key components such as state synchronization, real-time data exchange, and the integration of physical and virtual processes. Recent research has highlighted the growing importance of Digital Twins (DTs) in agriculture as a means to enable real-time monitoring, simulation, optimization, and automation of complex agro-food systems. Purcell [11] describes Digital Twins as high-fidelity virtual representations integrating continuous bidirectional data flows between physical and virtual entities. Their review demonstrates that most existing agricultural DT implementations remain limited to partially integrated systems, often referred to as Digital Shadows, where data primarily flows from the physical environment to the virtual model without continuous adaptive feedback mechanisms. Agricultural environments are inherently characterized by non-linearity, variability, and evolving operational conditions, making accurate modeling particularly challenging. Although data-driven approaches have demonstrated promising capabilities, current systems still face limitations

related to concept drift, non-stationarity, model generalization, and Sim-to-Real consistency. Furthermore, the integration of multimodal learning, distributed adaptation, and bidirectional Real-to-Sim/Sim-to-Real mechanisms remains largely underexplored in agricultural Digital Twin research. Recent agricultural DT reviews further classify applications across crop management, livestock monitoring, machinery optimization, resource management, and decision-support systems [12].

3. Digital Twin Systems and Industrial Applications

FCLA-DT [13] introduced a federated continual learning framework for distributed digital twin-based industrial IoT systems, integrating authentication mechanisms to ensure secure collaboration among multiple twins. Although the approach demonstrates strong predictive maintenance capabilities, its focus mainly remains on industrial fault diagnosis and secure collaboration. FCLLM-DT [14] introduced a Federated Continual Learning framework integrated with Digital Twins for distributed Industrial Internet of Things (IIoT) environments, targeting predictive maintenance and fault diagnosis applications. Related industrial DT research has also explored secure and self-optimizing architectures through blockchain-assisted Federated Learning [15]. Belay et al. [16] proposed a Digital Twin-based Federated Transfer Learning framework for anomaly detection in Industrial IoT systems, while preserving privacy. Emerging research explores the integration of digital technologies in agriculture and food systems. Xi et al. demonstrate the potential of augmented reality to support farm management tasks such as water quality monitoring and decision-making [17]. Similarly, Styliaras et al. present an AR-based application aimed at promoting sustainability awareness across the food lifecycle, from production to consumption [18]. These works illustrate the growing role of interactive and visualization technologies in the agro-food domain. However, they primarily focus on user interaction and decision support rather than adaptive or learning-based systems.

4. Adaptive and Distributed Learning

In dynamic environments, models must adapt over time without losing previously acquired knowledge. However, neural networks are prone to catastrophic forgetting, where learning new tasks leads to the loss of previously learned information. Kirkpatrick et al. propose Elastic Weight Consolidation (EWC) [19], enabling sequential learning while preserving prior knowledge, which is essential for real-world adaptive systems. Yu et al. [20] introduced a bidirectional Digital Twin and Mobile Edge Computing framework for Metaverse environments, enabling real-time synchronization and adaptive resource allocation through continuous interactions between physical and virtual entities. Kairouz et al. [21] reviewed Federated Learning (FL) as a distributed learning paradigm enabling collaborative model training across decentralized data sources while preserving data locality. Their study highlighted key challenges, including non-IID data distributions, communication overhead, and scalability limitations, which remain critical issues in large-scale distributed environments. The integration of FL and DTs has also been surveyed in smart-city contexts, emphasizing privacy-preserving local training and reduced raw-data transfer [22]. To this end, Wang et al. [23] proposed a Digital Twin (DT)-empowered federated incremental learning framework to address privacy preservation and performance degradation in highly heterogeneous, non-IID environments.

5. Summary

Although Digital Twins, multimodal sensing, Federated Learning, Continual Learning, and bidirectional physical-virtual interactions have been extensively investigated, these capabilities are typically studied in isolation or optimized toward specific objectives such as predictive maintenance, monitoring, fault diagnosis, resource allocation, or user interaction. Consequently, limited attention has been paid to continuously adaptive Digital Twin frameworks that jointly address multimodal process representation, distributed adaptation,

evolving operational conditions, and Sim-to-Real consistency in agro-food environments. This gap motivates the formulation presented in the following section.

III. PROBLEM FORMULATION - METHODOLOGICAL FRAMEWORK

The agro-food supply chain involves complex physical and operational processes characterized by non-linear dynamics, heterogeneous multi-modal data, and deformable material interactions. Accurately modeling such systems remains challenging due to the Sim-to-Real gap, i.e., the discrepancy between simulated models and real-world behavior, particularly in environments involving soft and deformable objects such as meat and curd.

Existing simulation approaches rely on simplified assumptions or static models that fail to capture real-world variability, leading to inaccurate predictions, suboptimal control strategies, and reduced training reliability. Also, existing multimodal digital twin systems primarily focus on educational interaction and adaptive sensory feedback. The proposed methodology is introduced progressively. In this work, we first formulate the physical agro-food system and its multimodal observations, then present a continuously adaptive Digital Twin representation, continual and federated learning mechanisms, and finally a unified optimization objective that combines all framework components.

1. System Modeling

Data Acquisition and Quality Enhancement:

The effectiveness of the proposed framework depends critically on the availability of high-quality, representative multimodal data. In practical agro-food environments, such data are obtained through continuous sensing pipelines integrating visual, environmental, and interaction-based measurements. To ensure robustness across varying operational conditions, the data acquisition process must capture diverse system states under different environmental, temporal, and operational configurations. This includes variations in lighting, material properties, sensor noise, and process dynamics. In addition, data-driven models benefit from enhanced data diversity through augmentation strategies. Generative approaches can be employed to synthesize realistic samples that preserve structural and physical characteristics, thereby improving model generalization across domains. The integration of continuous sensing and data augmentation mechanisms enables the framework to operate under realistic conditions, where data distributions evolve over time and across locations (Figure 1). This is particularly important for maintaining performance in non-stationary and heterogeneous agro-food environments. This data-centric perspective complements the proposed learning framework, in which continuous adaptation mechanisms operate on dynamically evolving multimodal inputs. We model the agro-food environment as a partially observable, non-stationary stochastic dynamical system. Let $s_t^{real} \in S$ denote the latent system state at discrete time step t, evolving according to unknown dynamics:

$$s_{t+1}^{real} \sim P(s_{t+1}^{real} \mid s_t^{real}, a_t) \tag{1}$$

where $a_t \in A$ represents control or interaction actions, including human interventions, robotic manipulation, or automated control signals, and $P(\cdot)$ denotes the (unknown) transition distribution. The latent state s_t^{real} represents a compact representation of the underlying physical process. It may encode environmental conditions, process characteristics, material properties, and interaction dynamics that cannot be directly observed from raw measurements. Such latent representations provide an abstraction layer that captures complex system behavior while reducing the dimensionality of heterogeneous sensory inputs. Temporal dependencies are particularly important in agro-food systems, since the evolution of historical processes influences current system states.

Variations in environmental conditions, process stages, and material transformations may accumulate over time and therefore require modeling that goes beyond isolated observations.

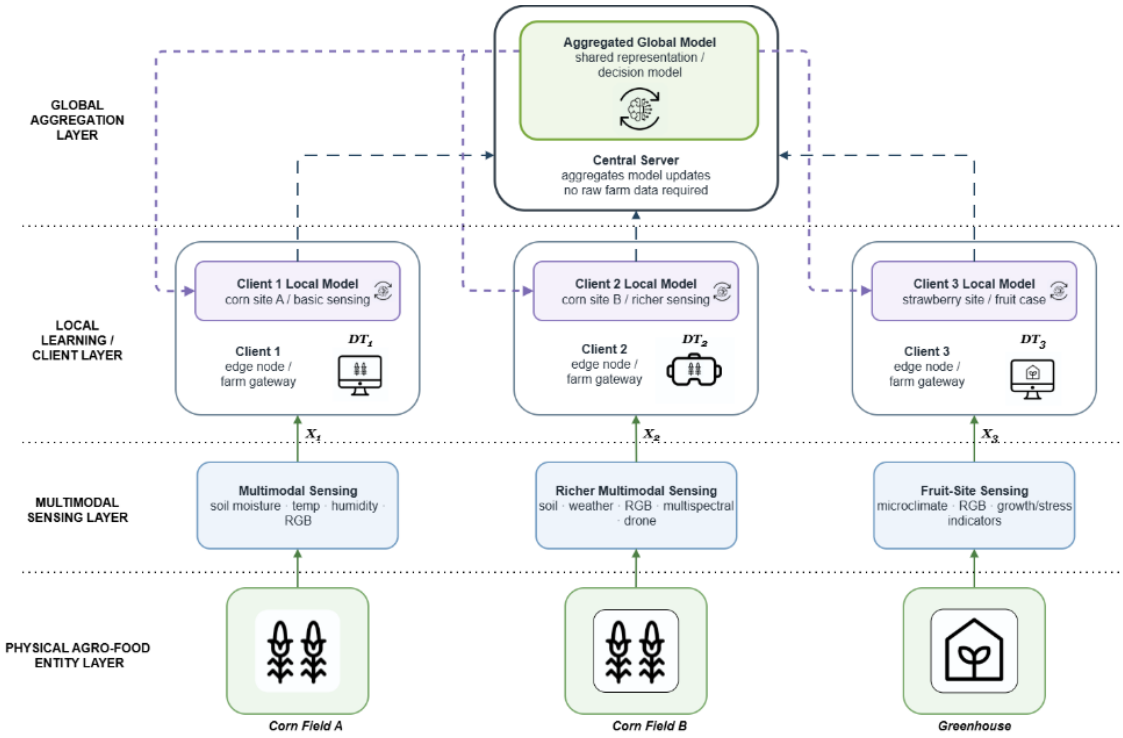


Figure 1. A depiction of a Federated Continual Learning Loop for Agro-Food Digital Twins. Agro-food physical entities (e.g., crop sites, greenhouses) provide multimodal data to local clients, reforming their Digital Twins; consequently, local updates are aggregated and continuously broadcast back. Furthermore, uncertainty naturally arises due to incomplete observations, noisy measurements, and stochastic process dynamics. Consequently, the transition process $P(\cdot)$ is treated probabilistically, allowing the framework to account for ambiguity and variability in real-world agro-food environments. This formulation enables modeling of both passive monitoring scenarios and active manipulation processes, which are particularly relevant in agro-food environments involving deformable materials. At each time step, the system produces a multi-modal observation vector composed of complementary sensing modalities:

$$x_t = \{x_t^{IoT}, x_t^{vision}, x_t^{force}\} \tag{2}$$

where individual modalities correspond to:

- x_t^{IoT}: denotes environmental and process measurements (e.g., temperature, humidity, chemical properties),
- x_t^{vision}: denotes visual observations (e.g., images, video streams),
- x_t^{force}: denotes interaction forces (e.g., robotic or haptic sensing).

Observations are generated through an implicit observation process:

$$x_t \sim O(x_t \mid s_t^{real}) \tag{3}$$

where $O(\cdot)$ denotes the observation process that maps the latent physical state to multimodal sensor measurements while accounting for partial observability, sensor noise, and modality-specific distortions.

2. Digital Twin Representation

The Digital Twin is modeled as a parameterized estimator that reconstructs the latent physical state from the accumulated multimodal observations.

$$\hat{s}_t = f_\theta(x_{1:t}) \tag{4}$$

where $\hat{s}_t$ denotes the estimated Digital Twin state, $x_{1:t}$ represents the sequence of multimodal observations collected from time step 1 to t, and f_θ is a parameterized model with learnable parameters θ. In this manner, the estimator continuously updates its internal representation as new observations become available, allowing the Digital Twin to capture both the current system state and its temporal progress under partially observable conditions. The continuously adaptive Digital Twin representation may also support XR-enabled interaction environments, allowing users to visualize and inspect multimodal process states through immersive interfaces. In the proposed framework, XR is not considered a core modeling component, but rather an interaction layer operating on top of the Digital Twin representation. Such environments may facilitate the interpretation of complex process behavior and support future applications involving operator training, monitoring, and decision support within agro-food ecosystems.

3. Sim-to-Real Discrepancy

Due to modeling approximations and incomplete knowledge of the system dynamics, a discrepancy arises between the real and simulated states. To quantify the consistency between the physical process and its Digital Twin representation, we define the Sim-to-Real discrepancy as:

$$\mathcal{L}_{sim2real} = \mathbb{E}_t[d(s_t^{real}, \hat{s}_t)] \tag{5}$$

where $d(\cdot,\cdot)$ is a task-dependent distance metric defined over relevant state variables, including geometric deformation, physical quantities, and process-specific indicators, while E_t represents the normalized temporal average value.

This formulation allows the Sim-to-Real gap to capture both perceptual discrepancies (due to sensing limitations) and dynamic discrepancies (due to inaccuracies in modeling system behavior). Minimizing $L_{sim2real}$ continuously aligns the Digital Twin representation with the evolving physical process.

4. Continuous Learning Requirement

Agro-food environments inherently exhibit non-stationary behavior. Changing environmental conditions and process variability induce temporal distribution shifts, which can be expressed as:

$$P_t(x, y) \neq P_{t+1}(x, y) \tag{6}$$

where $P_t(x, y)$ denotes the joint data distribution observed at time t.

This phenomenon, commonly referred to as concept drift, may progressively reduce model consistency and predictive performance if static learning strategies are employed [24]. As a result, digital twin representations should not be treated as fixed models but rather as continuously evolving entities that adapt to changing system conditions while preserving previously acquired knowledge [12].

To address this requirement, the Digital Twin parameters are updated incrementally using gradient-based optimization:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_t \tag{7}$$

where η denotes the learning rate and L_t the loss computed from the observations acquired at time t. The update mechanism is performed incrementally, enabling adaptation without requiring complete retraining from scratch. This process defines the Real-to-Sim adaptation loop, in which incoming real-world observations continuously refine the digital twin representation for site *i*. To mitigate catastrophic forgetting during continuous adaptation, the update is regularized as:

$$P_i(t) \rightarrow X_i(t) \rightarrow \theta_{i,t+1} \rightarrow DT_i(t+1) \tag{8}$$

However, naive incremental updates may cause catastrophic forgetting, where newly acquired information overwrites previously learned knowledge. To mitigate this issue, dynamic learning mechanisms may incorporate regularization strategies, replay-based memory mechanisms, or parameter stabilization approaches that preserve historical representations while enabling adaptation to emerging conditions [19].

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta [\mathcal{L}_t + \lambda \Omega(\theta_t, \theta_{t-1})] \tag{9}$$

where $\Omega(\cdot)$ is a regularization term preserving previously acquired knowledge and λ controls the trade-off between adaptation and knowledge retention.

5. Distributed Learning Setting

We now extend the formulation to a distributed learning setting consisting of N agro-food sites. Each site $i \in \{1, ..., N\}$ is associated with local datasets $D_i = \{(x_{i,k}, y_{i,k})\}_{k=1}^{n_i}$, where $x_{i,k}$ denotes a multimodal observation, $y_{i,k}$ the corresponding target variable, that may follow different (non-IID) distributions due to environmental and operational variability, and n_i the number of samples available at site *i*.

To leverage all available data without centralizing it, a federated learning setting is adopted. Each site learns local parameters θ_i that constitute the local learning objective, which is defined as:

$$\theta_i = \arg\min_\theta \mathcal{L}_i(\theta; \mathcal{D}_i) \tag{10}$$

where θ_i denotes the parameters of the local model and $L_i(\cdot)$ the local optimization objective computed using only dataset D_i, the locally updated models are periodically aggregated to obtain the global model:

$$\theta = \sum_{i=1}^{N} w_i \theta_i, \quad \sum_{i=1}^{N} w_i = 1 \tag{11}$$

where w_i denotes the aggregation weight assigned to site i and satisfies $\sum_i w_i = 1$.

$$w_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \tag{12}$$

This corresponds to a collaborative learning process across distributed environments, meaning raw data remain local and only focused model updates are exchanged for aggregation.

$$D_i \nrightarrow Server, \quad \Delta\theta_i \rightarrow Server \tag{13}$$

The non-IID nature of the data introduces additional challenges, as local models may exhibit biases toward specific conditions (e.g., crop types). In practice, distributed agro-food environments continuously generate evolving observations rather than static datasets. Consequently, local clients may experience different forms of concept drift, seasonal variations, and changing operational conditions, leading to distributed and time-varying data distribution:

$$\mathcal{D}_{i,t} \sim P_{i,t}(x,y) \quad P_{i,t}(x,y) \neq P_{j,t}(x,y) \quad P_{i,t}(x,y) \neq P_{i,t+1}(x,y)$$

where the first expression defines the local data distribution, the second captures spatial heterogeneity among different sites, and the third represents temporal concept drift at each site.

Therefore, the proposed framework extends conventional federated learning toward Federated Continual Learning (FCL), where:

$$\theta_{i,t+1} = \theta_{i,t} - \eta \nabla_\theta [\mathcal{L}_{i,t} + \lambda \Omega_i(\theta_{i,t}, \theta_{i,t-1})] \tag{14}$$

where $\Omega_i(\cdot)$ preserves previously acquired knowledge while adapting the model to newly observed data. The updated local models are then synchronized through periodic federated aggregation:

$$\theta_{t+1} = \sum_{i=1}^{N} w_i \theta_{i,t+1} \tag{15}$$

6. Problem Statement

The previous formulations describe the individual components of the proposed framework, including Digital Twin estimation, physical-virtual consistency, continual adaptation, and distributed learning. These components are jointly expressed through the following overall learning objective, which:

- minimizes the Sim-to-Real gap in agro-food systems,
- adapts to non-stationary and heterogeneous environments,
- integrates real-time multi-modal sensor data,
- supports accurate simulation and decision-making.

and is formulated into:

$$\min_{\theta_t} \mathbb{E}_{i,t}[\mathcal{L}_i(\theta_t; X_i(t)) + \lambda_1 d(P_i(t), DT_i(t)) + \lambda_2 \Omega(\theta_t, \theta_{t-1}) + \lambda_3 G_i(\theta_t)] \tag{16}$$

where $L_i(\theta_t; X_i(t))$ calculates the local learning from new agro-food observations, $\Omega(\theta_t, \theta_{t-1})$ shows the continual learning against the catastrophic forgetting while $d(P_i(t), DT_i(t))$ translates to the Real-to-Sim adaptation loop and $G_i(\theta_t)$ depicts the weighted generalization penalty for each site.

This formulation explicitly captures the interaction between sensing, modeling, and decision-making processes, treating the digital twin as an adaptive learning entity rather than a static representation.

7. Objective Function

The overall learning objective combines accuracy, stability, and generalization expressed as:

$$\min_{\theta} \mathbb{E}[\mathcal{L}_{task} + \lambda_1 \mathcal{L}_{sim2real} + \lambda_2 \mathcal{L}_{cont} + \lambda_3 \mathcal{L}_{gen}] \tag{17}$$

where:

- L_{task} denotes the task-specific learning objective optimized locally at each Digital Twin instance.
- $L_{sim2real}$ corresponds to the physical-virtual consistency term introduced in Equation (5).
- L_{cont} regularizes continual adaptation to mitigate catastrophic forgetting during incremental updates.
- L_{gen} promotes generalization across heterogeneous distributed environments.

This depiction (Figure 2) illustrates how the proposed framework integrates its main components within a unified learning architecture. The corresponding objective function formalizes the optimization criteria by synergistically balancing task performance, physical-virtual consistency, continual adaptation, and generalization across distributed agro-food environments.

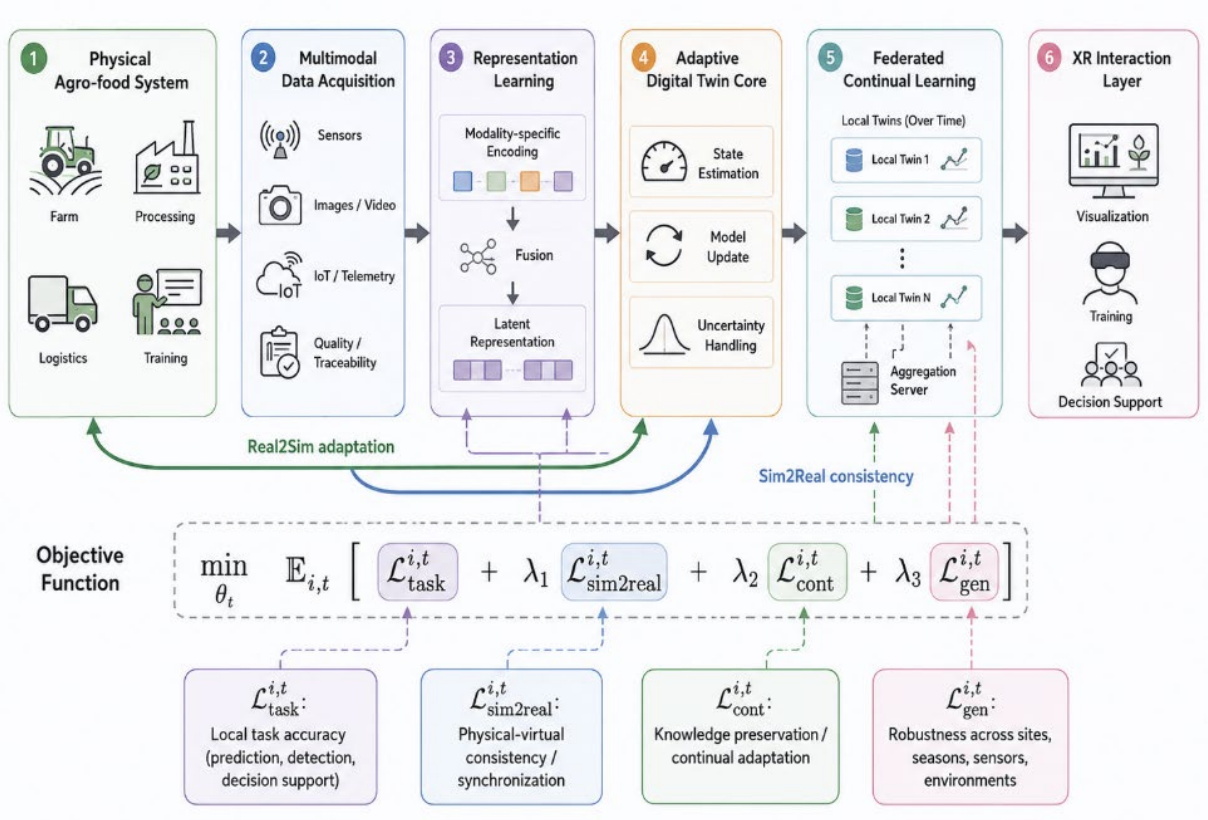


Figure 2. Simplified AI-XR Digital Twin Framework with Objective Function Mapping

IV. DISCUSSION AND COMPARATIVE ANALYSIS

The proposed framework extends existing Digital Twin (DT) paradigms beyond stationary or monitoring representations toward continuously adaptive multimodal learning systems that operate in dynamic agro-food environments. To this end, the primary contribution of this work is the formulation of a unified methodological framework describing how these complementary technologies can operate together within continuously adaptive Digital Twin ecosystems. As seen in the literature, DT-assisted Federated Learning frameworks employ Digital Twins as support entities for synthetic data generation or pretraining. FCLA-DT and FCLLM-DT introduced Federated Continual Learning mechanisms targeting predictive maintenance and industrial fault diagnosis while emphasizing privacy preservation and authentication. Although these systems support distributed adaptation, their Digital Twin representations remain primarily task-oriented and do not evolve according to changing physical process dynamics. Furthermore, recent bidirectional DT architectures in Metaverse environments have demonstrated how Digital Twins and edge computing can cooperate through feedback loops to optimize resources and enhance Quality-of-Experience. Nevertheless, such systems primarily optimize communication and computational resources rather than model the evolution of physical processes. In contrast, the proposed formulation investigates the coordinated integration of continuous Real-to-Sim adaptation via online multimodal observations; Sim-to-Real feedback for predictive and operational support; multimodal representation learning; Federated Continual Learning in evolving non-IID environments; support for deformable agro-food processes and hybrid physical dynamics; and XR-enabled human-in-the-loop interaction. Table 1 summarizes the conceptual differences between representative approaches and the proposed framework. The comparison demonstrates that several existing approaches incorporate individual components relevant to the proposed

framework, including Digital Twins, Federated Learning, Continual Learning, bidirectional interactions, and multimodal sensing. However, these capabilities are typically introduced to address isolated objectives such as predictive maintenance, fault diagnosis, communication optimization, or privacy preservation. In contrast, the proposed framework investigates their integration within a unified continuously adaptive Digital Twin architecture for agro-food systems, enabling multimodal representation learning, Real-to-Sim and Sim-to-Real adaptation, Federated Continual Learning under evolving non-IID conditions, and immersive XR-based interaction. Through this integration, the Digital Twin is no longer treated as a static monitoring or optimization entity, but as an evolving learning system that continuously adapts alongside the physical environment.

Table 1. Conceptual Comparison with Representative Digital Twin Frameworks

Method	**DT**	**Bidirectional**	**FCL**	**Multimodal**	**Sim-to-Real**	**XR**	**Agro-food**
Industrial DT approaches	Yes	No	No	No	No	No	No
Xia et al. (FCLA-DT)	Yes	No	Yes	No	No	No	No
Xia et al. (FCLLM-DT)	Yes	No	Yes	No	No	No	No
Yu et al.	Yes	Yes	No	No	Partial	Yes	No
Verdouw et al.	Yes	Partial	No	Partial	No	No	Yes
Belay et al.	Yes	No	No	No	Partial	No	No
Wang et al.	Yes	No	Partial	No	No	No	No
Proposed Framework	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Beyond the architectural integration presented in this work, the proposed framework establishes a methodological basis for developing next-generation adaptive Digital Twin ecosystems for agro-food applications. The proposed formulation enables the continuous interaction between multimodal sensing, adaptive Digital Twin representations, distributed learning, and immersive user interfaces. This integrated perspective supports future applications involving precision agriculture, greenhouse monitoring, robotic manipulation of deformable food products, and operator training through XR-assisted environments.

1. Implementation Perspective

The proposed framework is currently being implemented through representative agro-food scenarios that involve distributed multimodal sensing and continuously adaptive Digital Twin synchronization. Building on our previous work on XR applications for food science [25], the current implementation focuses on establishing the interaction among multimodal sensing, adaptive Digital Twin synchronization, Federated Continual Learning, and immersive XR environments within a unified adaptive ecosystem. At this stage, the contribution of this work is summarized in the conceptual architecture and methodological formulation of the proposed framework, while implementation and quantitative evaluation are in progress. Future experimental studies will validate the proposed methodology through representative agro-food scenarios, evaluating Digital Twin consistency, Sim-

to-Real fidelity, adaptation under concept drift, distributed learning performance, scalability, and decision-support effectiveness.

V. CONCLUSION

This paper presented a conceptual and methodological framework for continuously adaptive multimodal Digital Twins in agro-food environments. Motivated by the limitations of existing Digital Twins in presenting adaptation capabilities, the proposed formulation redefines Digital Twins as dynamic ecosystems. The proposed framework integrates multimodal sensing, bidirectional Real-to-Sim and Sim-to-Real adaptation, Federated Continual Learning, and XR-assisted interaction within a unified architecture that supports learning and adaptation under dynamic, heterogeneous, and non-stationary agro-food conditions. Future work will focus on implementing and experimentally validating the proposed framework across representative agro-food scenarios, extending our previous research on XR technologies for food science toward continuously adaptive Digital Twin ecosystems.

REFERENCES

[1]. F. Muratore, F. Ramos, G. Turk, W. Yu, M. Gienger, and J. Peters, "Robot learning from randomized simulations: A review," Frontiers in Robotics and AI, vol. 9, p. 799893, 2022.

[2]. J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 23–30.

[3]. X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 3803–3810.

[4]. J. Tremblay et al., "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 969–977.

[5]. A. Pitkevich and I. Makarov, "A survey on sim-to-real transfer methods for robotic manipulation," in 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY), 2024, pp. 000259–000266. DOI: 10.1109/SISY62279.2024.10737545.

[6]. L. Da, J. Turnau, T. P. Kutralingam, A. Velasquez, P. Shakarian, and H. Wei, "A survey of sim-to-real methods in RL: Progress, prospects and challenges with foundation models," arXiv preprint arXiv:2502.13187, 2025.

[7]. Y. Jang, J. Baek, S. Jeon, and S. Han, "Bridging the simulation-to-real gap of depth images for deep reinforcement learning," Expert Systems with Applications, vol. 253, p. 124310, 2024.

[8]. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2018.

[9]. Y. Nomura and H. Mineno, "Toward multimodal agricultural digital twins: Integrating autonomous UGV monitoring, high-precision phenotyping, and diffusion-based generative data augmentation," in The First International Workshop on AI in Agriculture, 2026. [Online]. Available: https://openreview.net/forum?id=kytnCzrBvD.

[10]. D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterizing the digital twin: A systematic literature review," CIRP Journal of Manufacturing Science and Technology, vol. 29, pp. 36–52, 2020.

[11]. W. Purcell and T. Neubauer, "Digital twins in agriculture: A state-of-the-art review," Smart Agricultural Technology, vol. 3, p. 100094, 2023.

[12]. R. Zhang, H. Zhu, Q. Chang, and Q. Mao, "A comprehensive review of digital twins technology in agriculture," Agriculture, vol. 15, no. 9, p. 903, 2025.
[13]. Y. Xia, X. Liu, Y. Zhao, and Y. Wang, "FCLA-DT: Federated continual learning with authentication for distributed digital twin-based industrial IoT," Journal of Communications and Information Networks, vol. 9, no. 4, pp. 362–373, 2024.
[14]. Y. Xia et al., "FCLLM-DT: Enpowering federated continual learning with large language models for digital-twin-based industrial IoT," IEEE Internet of Things Journal, vol. 12, no. 6, pp. 6070–6081, 2024.
[15]. I. B. Ababio et al., "A blockchain-assisted federated learning framework for secure and self-optimizing digital twins in industrial IoT," Future Internet, vol. 17, no. 1, 2025, ISSN: 1999-5903. [Online]. Available: https://www.mdpi.com/1999-5903/17/1/13.
[16]. M. A. Belay, A. Rasheed, and P. S. Rossi, "Digital twin-based federated transfer learning for anomaly detection in Industrial IoT," in 2025 IEEE Symposium on Computational Intelligence on Engineering/Cyber Physical Systems (CIES), IEEE, 2025, pp. 1–7.
[17]. M. Xi, M. Adcock, and J. McCulloch, "Future agriculture farm management using augmented reality," in 2018 IEEE Workshop on Augmented and Virtual Realities for Good (VAR4Good), 2018, pp. 1–3. DOI: 10.1109/VAR4GOOD.2018.8576887.
[18]. G. Styliaras, S. Tzima, and V. Dimou, "AR-based sustainability in food development," in 2022 13th International Conference on Information, Intelligence, Systems Applications (IISA), 2022, pp. 1–8. DOI: 10.1109/IISA56318.2022.9904381.
[19]. J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," Proceedings of the National Academy of Sciences, vol. 114, no. 13, pp. 3521–3526, 2017.
[20]. J. Yu, A. Alhilal, P. Hui, and D. H. Tsang, "Bi-directional digital twin and edge computing in the metaverse," IEEE Internet of Things Magazine, vol. 7, no. 3, pp. 106–112, 2024.
[21]. P. Kairouz and H. B. McMahan, "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, no. 1-2, pp. 1–210, 2021.
[22]. S. P. Ramu et al., "Federated learning enabled digital twins for smart cities: Concepts, recent advances, and future directions," Sustainable Cities and Society, vol. 79, p. 103663, 2022.
[23]. Q. Wang, S. Chen, M. Wu, and X. Li, "Digital twin-empowered federated incremental learning for non-IID privacy data," IEEE Transactions on Mobile Computing, vol. 24, no. 5, pp. 3860–3877, 2024.
[24]. M. Savoia, D. Annunziata, D. Thakur, G. Fortino, and F. Piccialli, "Federated continual learning meets digital twins: A survey on methods, intersections and perspectives," Neurocomputing, p. 133366, 2026.
[25]. G. Protogeros et al., "Conceptualizing and advancing on extended reality applications in food science and technology," Journal of Food Engineering, vol. 396, p. 112557, Mar. 2025. DOI: 10.1016/j.jfoodeng.2025.112557.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Water Quality Monitoring and Pollution Detection Using Edge Internet of Things-Based Hybrid Artificial Intelligence Models: A Systematic Literature Review

Hüseyin Özgür KAZAN
Graduate School of Natural and Applied Sciences, Gazi University
Ankara, Turkey
25833304009@gazi.edu.tr
ORCID: 0000-0001-6812-6927

İsmail ATACAK
Department of Computer Engineering, Faculty of Technology, Gazi University,
Ankara, Turkey
iatacak@gazi.edu.tr,
ORCID: 0000-0002-6357-0073

İbrahim Alper DOĞRU
Department of Computer Engineering, Faculty of Technology, Gazi University
Ankara, Turkey
iadogru@gazi.edu.tr
ORCID: 0000-0001-9324-7157

Abstract

This study presents a systematic review of research on water quality monitoring and water pollution detection using a hybrid artificial intelligence (AI) system developed for Edge-IoT technology, published in the Web of Science Core Collection from 2023 to 2026, indexed in the SCI-EXPANDED and ESCI. Three complementary search strategies were used: a main Edge-IoT and artificial intelligence (AI) search, a secondary search based on metaheuristic optimization, and a supplementary search based on large language models (LLMs), generative artificial intelligence (GenAI), and retrieval-augmented generation (RAG). A total of 68 primary search records were found. After one duplicate record was removed, 67 unique candidate studies were evaluated. The selection flow followed the PRISMA guidelines, and the candidate studies were assessed using a five-item eligibility matrix, resulting in 37 studies in the final analysis set. The findings show that low-cost edge hardware, such as Raspberry Pi, ESP32, Arduino, and Jetson/Orin, is widely deployed in field applications. Additionally, it has been observed that machine learning, deep learning, computer vision, ensemble learning, metaheuristic optimization, digital twins, blockchain, and LLM/GenAI/RAG-based decision-support methods are complementary across various problem areas. This study systematically identifies current trends and research gaps for real-time, explainable, low-latency, and field-deployable water quality decision-support systems.

Keywords: water quality monitoring, water pollution detection, Edge-IoT, edge computing, hybrid artificial intelligence

I. INTRODUCTION

The safe monitoring of water resources is a critical research topic for drinking water security, river basin management, and aquaculture systems [1-8]. In wastewater treatment plants, water treatment processes, and environmental pollution scenarios, real-time monitoring and early warning approaches are becoming increasingly important [9-15]. Although conventional laboratory-based or centralized measurement approaches can provide high accuracy, they have limitations in terms of real-time monitoring, early warning generation, low-cost field deployment, and continuous observation over large areas [2], [4-7], [16]. Therefore, faster, local, and intelligent monitoring approaches are needed to track dynamic changes in parameters such as pH, dissolved oxygen, turbidity, total dissolved solids, nitrate/nitrite, heavy metals, and related water quality indicators [4], [6-8], [17-21]. Edge-Internet of Things (IoT) and Artificial Intelligence of Things (AIoT) architectures enable water quality data to be collected directly in the field, processed locally on edge devices, and, when necessary, used to support on-device decision-making mechanisms [2-5], [10],[11], [16], [21-25]. Edge platforms such as Raspberry Pi, ESP32, Arduino, and Jetson Orin address different application requirements in terms of cost,

energy consumption, latency, and computational capacity [4], [5], [10], [11], [16], [21-26]. Machine learning, deep learning, computer vision, federated learning, digital twin, and metaheuristic optimization approaches are used for sensor data interpretation, anomaly detection, water quality prediction, visual pollution detection, and decision-support processes [6-14], [17-21], [23-35]. More recently, Large Language Model (LLM), Generative Artificial Intelligence (GenAI), and Retrieval-Augmented Generation (RAG)-based approaches have introduced new opportunities for interpreting water quality data, modeling sensor noise, querying data through natural language, and supporting decision-making PROCESSES [15], [36],[37]. This study aims to systematically review recent studies published between 2023 and 2026 and present the current state of Edge-IoT-based hybrid artificial intelligence approaches for water quality monitoring and pollution detection. The studies included in the final analysis set were evaluated with respect to AI models running on edge devices, vision-based detection systems, metaheuristic optimization, federated learning/digital twin architectures, and LLM/GenAI/RAG-based decision-support solutions.

The main contributions of this study can be summarized as follows:

- A recent and focused pool of Edge-IoT-based water quality studies covering the 2023–2026 period was constructed.
- Edge-IoT, metaheuristic optimization, federated learning/digital twin, and LLM/GenAI/RAG approaches were examined from an integrated perspective.
- A transparent and reproducible selection process was applied, and studies satisfying the eligibility criteria were included in the final analysis set without imposing a subjective restriction.
- The studies were synthesized in terms of thematic category, hardware, method, performance metrics, and research gaps.

The remainder of this paper is organized as follows. The Background section presents the key concepts underlying the study. The Methodology section describes the systematic review process. The Findings and Thematic Analysis section provides the thematic synthesis of the selected studies. The Discussion section evaluates the results in relation to the research questions, and the Conclusion section summarizes the main findings and outlines future research directions.

II. BACKGROUND

This section summarizes the key concepts that define the scope of this study, including water quality monitoring, water pollution detection, Edge-IoT/AIoT, hybrid artificial intelligence, metaheuristic optimization, and LLM/GenAI/RAG-based decision support.

1. Water Quality Monitoring and Water Pollution Detection

Water quality monitoring involves the regular measurement and interpretation of physical and chemical parameters, including pH, dissolved oxygen, temperature, total dissolved solids, turbidity, nitrate/nitrite, heavy metals, and related indicators [4–8], [17–21]. These parameters are used as key indicators for drinking water safety, river basin management, aquaculture systems, wastewater treatment processes, and environmental early warning applications [3–8], [12–15]. Water pollution detection, however, is not limited to numerical water quality parameters alone; floating debris, oil films, aquatic plant invasions, heavy metal accumulation, increased turbidity, and visual ecosystem indicators are also considered within this scope [9–11], [17],[18], [21], [23–25].

2. Edge-IoT- and AIoT-Based Monitoring

Edge-IoT architectures are distributed computing approaches that enable data collected from sensors to be processed on edge devices located close to the field before being transmitted to a central cloud. These architectures are becoming increasingly important in water quality monitoring and pollution detection applications, where latency, communication overhead, energy consumption, and real-time decision-making are critical [2–5], [10], [11], [16], [21–25]. AIoT combines IoT infrastructure with artificial intelligence models, enabling sensor data to be analyzed locally and used in decision-support processes [3],[4], [16]. Platforms such as Raspberry Pi, ESP32, Arduino, Jetson Nano, and Orin Nano address different application requirements in terms of computational capacity, energy consumption, and cost [4],[5], [10],[11], [16], [19], [21–26].

3. Hybrid AI, Optimization, and LLM-Based Decision Support

Hybrid artificial intelligence models are those that combine ML/DL models with optimization models, sensor network architectures, edge cloud integration, or decision-support models [14], [26-32], [35]. Metaheuristic algorithms such as Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Whale Optimization Algorithm (WOA), Grey Wolf Optimizer (GWO), Salp Swarm Optimization (SSO), Simulated Annealing (SA), and Knowledge-Based Multi-Objective Particle Swarm Optimization (KBMOPSO) are used for tasks including model hyperparameter tuning, sensor placement, missing data imputation, water quality/pollutant prediction, and system optimization [14], [26-31], [35]. More recently, LLM-, GenAI-, and RAG-based approaches have created new opportunities for interpreting sensor data, modeling sensor noise, generating expert-system-like recommendations, and enabling users to query data through natural language. Nevertheless, further research is still needed to assess the feasibility of these approaches on edge devices in terms of energy consumption, latency, reliability, explainability, and data privacy [15], [36],[37].

III. METHODOLOGY

This study adopts a systematic literature review approach. During the planning phase, the research scope, research questions, database, search strategies, inclusion and exclusion criteria, and quality assessment process were defined in advance. The analytical objective, scope, viewpoint, and context of the study are summarized in Table 1 in accordance with the Goal Question Metric (GQM) approach. Based on this framework, search pools were constructed, candidate studies were screened at the title and abstract levels, and full-text assessments were conducted for studies whose topical relevance or thematic classification could not be determined from the initial screening. Subsequently, eligibility scoring was applied, and the final thematic analysis set was synthesized.

Table 1. GQM-based Research Objective and Scope

GQM Component	**Description**
Object of analysis	Edge-IoT-based water quality monitoring, water pollution detection, wastewater monitoring, and hybrid AI studies
Purpose	To characterize, classify, compare, and identify research gaps in recent approaches
Scope	Edge-IoT/AIoT architectures, edge devices, AI methods, metaheuristic optimization, LLM/GenAI/RAG, and performance outcomes
Viewpoint	Researcher and systematic literature review perspective
Context	Article-type studies published between 2023 and 2026 in the WoS Core Collection SCI-EXPANDED and ESCI indices

In line with the scope and purpose defined in Table 1, the following research questions were formulated:

- RQ1: What are the current Edge-IoT and AIoT architectures used in water quality monitoring and water pollution detection?
- RQ2: Which ML, DL, computer vision, metaheuristic optimization, and LLM/GenAI/RAG-based method families are used in these studies?
- RQ3: What performance metrics are reported in the reviewed studies?
- RQ4: What are the main limitations and research gaps highlighted in the literature?

1. Review Planning Phase

During the planning phase, the topic domain, year range, index coverage, document type, and selection criteria were defined. The review scope was limited to studies that are directly related to water quality monitoring or water pollution detection, include an IoT/Edge-IoT/AIoT or edge-device component, and present an AI-, optimization-, federated learning-, digital twin-, or LLM/GenAI/RAG-based method. This scope allowed the study to be distinguished from the broader environmental monitoring literature and to focus specifically on Edge-IoT-based intelligent water quality systems.

2. Search Strategy

The search was conducted using filters that include the indices in which the article is indexed (SCI-EXPANDED and ESCI in the Web of Science Core Collection), document type, and year range (2023-2026). The search strategy was based on three complementary queries. The first query (Q1) focused on Edge-IoT and artificial intelligence-based water quality monitoring studies, whereas the second query (Q2) focused on metaheuristic optimization and hybrid artificial intelligence studies. The third (Q3) has focused on studies of LLMs, GenAI, RAG, and decision support. The complete search strings are presented in Table 2. The queries were designed using the Web of Science Topic Search logic via the TS field, followed by applying year, index, and document type filters.

Table 2. Web of Science Search Strings and Objectives

Query	**Search string**	**Objective**
Q1	TS=(("water quality" OR "water pollution" OR "wastewater" OR "drinking water" OR "river water" OR aquaculture OR aquaponic) AND ("Internet of Things" OR IoT OR "Edge-IoT" OR "edge computing" OR AIoT OR "smart sensor*" OR "wireless sensor network*" OR WSN) AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "computer vision" OR CNN OR YOLO OR LSTM OR "random forest" OR XGBoost))	To construct the main Edge-IoT, sensor, and AI study pool.
Q2	TS=(("water quality" OR "water pollution" OR wastewater OR aquaculture OR aquaponic OR "water treatment") AND (IoT OR "edge computing" OR "wireless sensor network*" OR sensor*) AND (metaheuristic OR optimization OR "particle swarm optimization" OR PSO OR "genetic algorithm" OR GA OR "grey wolf optimizer" OR GWO OR "whale optimization algorithm" OR WOA OR "salp swarm optimization" OR SSO OR "simulated annealing" OR SA))	To identify metaheuristic optimization and hybrid AI studies

Query	Search string	Objective
Q3	TS=(("water quality" OR wastewater OR "water pollution" OR aquaculture OR "water treatment") AND (IoT OR sensor* OR "edge computing" OR "digital twin") AND ("large language model*" OR LLM OR "generative AI" OR GenAI OR ChatGPT OR "retrieval augmented generation" OR RAG OR "decision support" OR "expert system"))	To identify LLM, GenAI, RAG, digital twin, and decision-support studies

The searches yielded 47 candidate records in the Q1 Edge-IoT/AI pool, 13 in the metaheuristic optimization candidate pool, and 8 in the LLM/GenAI/RAG candidate pool. In total, 68 candidate records were retrieved. After duplicate checking (based on DOI and title), 67 candidate studies were assessed, and 37 were included in the final analysis set based on eligibility scoring. This selection is summarized in Figure 1. The presence of only one duplicate record across the three query pools indicates that Edge-IoT/AI-, metaheuristic optimization-, and LLM/GenAI/RAG-oriented studies form largely complementary but only partially overlapping literature clusters in this domain.

3. Data Extraction Protocol

A standardized data extraction protocol was used when evaluating the candidate studies. This protocol was designed to capture not only bibliographic information but also the application context, monitored water-quality parameters or target outputs, data types, hardware and architectural components, AI/optimization methods used, reported performance metrics, types of evidence, reported limitations, and thematic categories. In this way, the studies were made comparable in terms of methods, architectures, data types, and performance. During the data extraction process, model-, system-, and evidence-level metrics were recorded, including accuracy, precision, recall, specificity, F1-score, Area Under the Curve (AUC), False Positive Rate/False Negative Rate (FPR/FNR), coefficient of determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), mean Average Precision (mAP), Average Precision (AP), Dice coefficient, Intersection over Union (IoU), Topographic Error (TE), Quantization Error (QE), Silhouette Index, Dunn Index, Signal-to-Noise Ratio (SNR), recovery, Relative Standard Deviation (RSD), detection limit, response time, compression ratio, latency/delay, Frames Per Second (FPS), Central Processing Unit (CPU)/memory/power usage, model size, Giga Floating-Point Operations per Second (GFLOPs), battery life, bandwidth/data transfer reduction, blockchain gas/transaction cost, throughput, trust score, fault tolerance, BERTScore, FactCC, ROUGE-L, Question Answering (QA) accuracy, Precision@1/Recall@5, decision error, and user/expert feedback. These fields served as the basis for the synthesis matrix presented in Table 4 and for generating distributions of themes, hardware, methods, and performance metrics.

4. Inclusion and Exclusion Criteria

The following criteria were used to choose the relevant studies: the study had to be published between 2023 and 2026 as an Article in SCI-EXPANDED or ESCI; the study had to be directly related to water quality monitoring or water pollution detection; the study had to contain an IoT/Edge-IoT/AIoT, sensor network, or edge-device component; the study had to present an ML-, DL-, optimization-, federated learning-, digital twin-, or LLM/GenAI/RAG-based method or decision-support approach. Exclusion criteria were: out of year, out of index, or out of document type, document had weak relevance to water quality, did not have an IoT/sensor/edge component, did not present an intelligent method or decision-support component, and only provided conceptual background material.

5. Quality Assessment and Eligibility Scoring

Candidate studies were evaluated using a five-item eligibility matrix. Each criterion was assigned a score of 1 or 0, and each study received a total eligibility score ranging from 0 to 5. The criteria and decision thresholds are presented in Table 3. As a result of this process, 22 studies were classified as "Definite Include", 15 studies as "Include", and a total of 37 studies were included in the final analysis set.

Table 3. Eligibility Assessment Criteria and Decision Thresholds

Code/Score	**Assessment Description**	**Scoring/Decision**
Q1	Direct relevance to water quality monitoring or water pollution detection	Yes=1, No=0
Q2	Inclusion of IoT, Edge-IoT, AIoT, sensor network, or edge-device component	Yes=1, No=0
Q3	Inclusion of an ML-, DL-, optimization-, federated learning-, digital twin-, or LLM/GenAI/RAG-based intelligent method or decision-support component	Yes=1, No=0
Q4	Reporting of a performance metric or measurable evaluation outcome	Yes=1, No=0
Q5	Inclusion of real data, field deployment, prototype, simulation, benchmark, or case study	Yes=1, No=0
5	Studies satisfying all eligibility criteria	Definite Include
4	Studies that miss one criterion but are strongly aligned with the research scope.	Include
3	Studies are partially related to the scope or limited in evidence/method components.	Reserve/Borderline
0–2	Studies that are weakly related to the scope or do not satisfy the core criteria	Exclude/Background

For the 15 studies classified as "Include", the missing criterion was most commonly associated with limited reporting of measurable system-level metrics or a narrower field/prototype validation scope. However, these studies were retained in the final synthesis because their methodological contributions and topical relevance strongly aligned with the review's scope.

6. Review Phase

During the review phase, candidate records were processed through duplicate checking, title/abstract screening, full-text assessment when necessary, eligibility scoring, and thematic classification. Studies classified as Reserve/Borderline, where topical relevance or thematic classification was unclear, were checked through full-text review. Studies classified as Definite Include or Include were retained in the final analysis set, while the remaining studies were excluded from the analysis. The process is summarized in Figure 1.

7. PRISMA-Like Study Selection Flow

The study selection process was summarized using a PRISMA-like transparent reporting structure. Records retrieved from the three search pools were transformed into the final analysis set after duplicate removal, title/abstract screening, full-text assessment when required, and eligibility scoring. The process is presented in Figure 1.

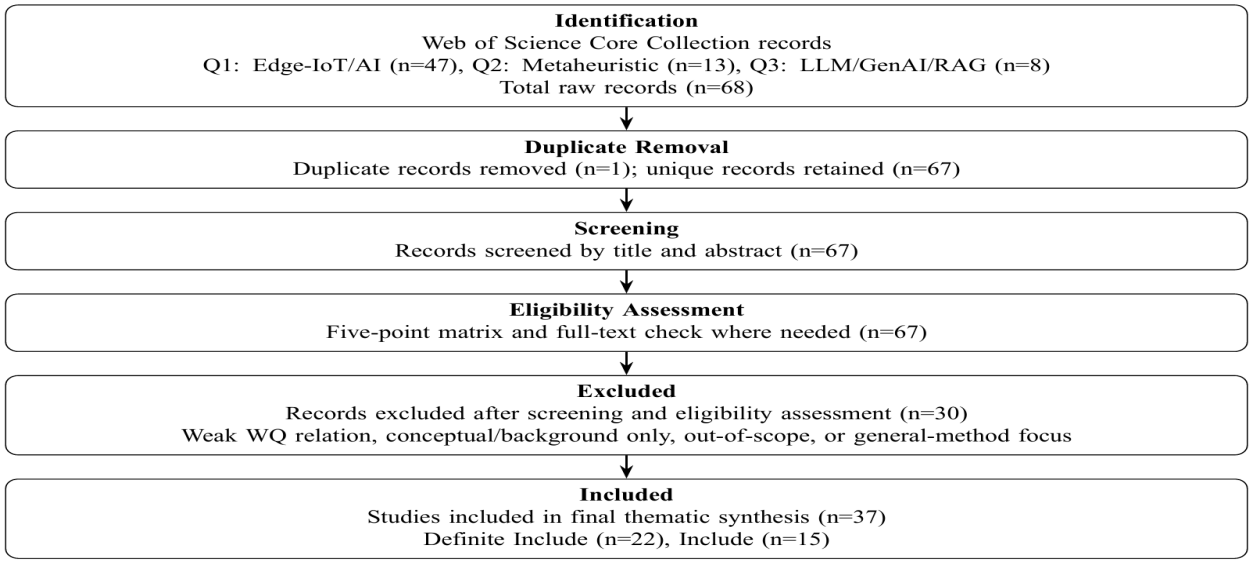


Figure 1. PRISMA-like study selection flow.

IV. FINDINGS AND THEMATIC ANALYSIS

This section presents a thematic analysis of the 37 studies included in the final analysis set. The synthesis matrix in Table 4 summarizes each study by reference number, thematic group, application domain, method/hardware used, and the reported metric or type of evidence. The findings were evaluated from multiple perspectives, including publication year, theme, method, hardware, performance metric, and application context. This evaluation shows that the water quality monitoring literature is not limited to sensor-based measurement systems; it has expanded to include vision-based detection, edge-cloud integration, federated learning, digital twins, metaheuristic optimization, and LLM/GenAI/RAG-based decision-support approaches. The theme codes used in Table 4 are as follows: T1 = Edge-IoT monitoring; T2 = WQ prediction/modeling; T3 = vision-based detection; T4 = metaheuristic/hybrid AI; T5 = digital twin/federated learning/edge-cloud/blockchain; T6 = LLM/GenAI/RAG/decision support. The abbreviations in the table are defined as follows: Acc., Prec., Rec., BW, comp., err., TE, QE, Silh., NL, and WQ refer to accuracy, precision, recall, bandwidth, compression, error, topographic error, quantization error, Silhouette, natural language, and water quality, respectively. The abbreviations in the table are consistently used throughout.

Table 4. Synthesis Matrix of the Final Analysis Set

Ref.	Theme	Year	Application	Method	Metric
[1]	T1	2023	CIS-WQMS	Edge devices, ensemble ML/SSL	Acc. 99.46%; F_1 99.27%; 4.44 ms; AUC 1.00
[2]	T1	2023	IoTEC environmental monitoring	IoTEC, edge ML, regression/image processing	Acc. 87% (reg.), > 90% (image); latency -13%; data -50%; battery +130%
[3]	T1	2024	AIoT aquaculture inspection	ESP32, AIoT, SRU	MAE/RMSE/MAPE; SRU > LSTM

Ref.	Theme	Year	Application	Method	Metric
[4]	T1	2025	Tilapia aquaculture prediction	ESP32, MLR/DTR/RFR	R^2 0.81–0.83; RMSE/MAE; 0.32 kWh/cycle
[5]	T1	2025	Household drinking water monitoring	Raspberry Pi, fuzzy logic	Acc. 98.724%
[22]	T1	2026	Turbidity estimation in farm ponds	Raspberry Pi 4, CNN	Acc. 96.3%; 46 ms; CPU 35–40%
[16]	T1	2026	Low-cost on-device WQ monitoring	ESP32, TinyML	Acc. 99.28%; Prec./Rec. 0.99; AUC 1.00; cost <$80
[6]	T2	2025	River WQ monitoring	Raspberry Pi 5, MRMX-WQM	Acc. 99.8%/100%; err. cost 2/0; 5k/59k obs/s
[17]	T2	2025	Heavy metal sensing	Edge biosensor, ML	RMSE 0.032 nM; AUC >0.99 ; 28 ms; 2.3 W; recovery 92–105%
[7]	T2	2025	Cauvery Basin WQ monitoring	Raspberry Pi 5, MLRMC-WQM	Acc. 97.92%; seasonal Acc. 99.8%/100%; err. cost 2/0
[8]	T2	2025	Macao WQ dataset	Raspberry Pi, RF/IF/LR/LOF	F_1 98.99%; Acc. 98.10%; AUC 0.91
[18]	T2	2025	CQD biosensor contaminants	CNN-LSTM attention	R^2 0.968; F_1 0.944; 0.31 s; LOD 5–18 nM
[19]	T2	2025	DO classification	Raspberry Pi, XGBoost, SMOTE	Acc. 96.6%; macro AUC 0.907
[20]	T2	2026	Sparse coding WQ time series	BASM, sparse coding, ESN comparison	AUC 0.87–0.93; MAE 0.0342; comp. 27.4:1; <50 ms
[9]	T3	2023	Urban sewage sludge detection	Mobile edge, Faster R-CNN, CNN split deployment	Delay 94.20 ms; CPU growth 6.50
[10]	T3	2023	Floating debris detection	Raspberry Pi 4B, TC-YOLOv5	mAP 84.2%; 1.18 s; 3.7 GFLOPs; 3.6 MB
[11]	T3	2025	Oil spill measurement	Raspberry Pi 4B, YOLOv5n, TensorRT	Acc. 96.6%; mAP 88.2%; 13.6 FPS; 4.2 W
[23]	T3	2026	Pond water visual assessment	Raspberry Pi 2W, ANN vision	Acc. > 99%; confusion matrix
[24]	T3	2026	Fish detection/classification	Jetson Nano, YOLOv4-Tiny, IoT dashboard	mAP 97.14%; AP/ $F_1 \approx 1.00$; 8 FPS; 23 MB
[21]	T3	2026	Nitrate/nitrite vision detection	Orin Nano, VGG16, KNN/SVM	Acc. 99.92%/98.04%; CV 99.79%/98.07%; 404/359 ms
[25]	T3	2026	Water hyacinth detection	Jetson Nano, UNet/VGG19	Dice 0.927; IoU 0.875; Acc. 96%; AUC 0.994

Ref.	Theme	Year	Application	Method	Metric
[26]	T4	2024	Optimized NN WQ monitoring	Arduino, GA-PSO-ANN	Acc. 91%; Prec. 0.89; Rec. 0.94
[27]	T4	2024	Sewer sensor placement	ML-based SI, GA optimization	Real sewer case; observability/reliability scores
[28]	T4	2025	Heavy metal prediction	WGCBM, WOA-GWO, CNN-BiLSTM-Attention	Acc. 97.1%; Rec. 0.918; F_1 0.913; mem. 4.5 GB
[29]	T4	2025	Missing data in WQ networks	ABL tensor factorization, PSO adaptation	RMSE 0.2709–0.3068; MAE 0.1473–0.1739; rank 1.0
[30]	T4	2026	Distributed environmental clustering	DiSOM, DSOM, DPSO, DK-means	TE -3.25%; QE -30%; SI/DI; outlier Acc. 75%
[31]	T4	2026	Smart aquaponics optimization	ESP32, IoT sensors, GA/SA/PSO	Optimization-based aquaponics control
[32]	T5	2024	Precision aquaculture edge-cloud	Raspberry Pi 3B+, edge-cloud FL, YOLOv7	mAP 0.987; AR 0.969; Prec. >0.95 ; 2-farm validation
[12]	T5	2024	WTP digital twin platform	Microservices, LSTM, ASM1, Digital Twin	Prec. 0.984; Rec. 0.977; F_1 0.981
[13]	T5	2025	SW3T-IoT	Three-tier IoT, continual learning, CNN-BiLSTM-Attention	MAE 0.087–0.089; RMSE 0.121–0.124; MSE 0.0146–0.0154
[33]	T5	2025	Crayfish health assessment	YOLOv11, Raspberry Pi 5, federated learning	Acc. 87.4%; mAP 89.5%; 1.8 s; BW -90%; 3.2 W
[34]	T5	2025	Federated WQ management	FDL, IoT network	Acc. 98%; F_1 0.96; R^2 0.975; RMSE 18.19
[14]	T5	2025	ML-DT for MBR WWTPs	ML-DT, KBMOPSO, LSTM encoder-decoder	MAPE <6.45% ; MAE <3.71 ; R^2>0.87 ; aeration 0.06–0.12 kWh/t
[35]	T5	2026	Secure WQ monitoring	MLP-GRU, SSO, blockchain-smart contracts	Acc. 98.17%; AUC 0.983; 87 ms; false alarms -95.2%
[36]	T6	2025	RAG-LLM aquaculture control	RAG-LLM, DQN, IoT automation	Fish growth +1.8%; decision error 1.8%; QA Acc. 72%
[15]	T6	2025	GenAI wastewater sensor denoising	VAE, RF, ISM sensors	SNR 1.32–3.92; acc. +215.9%; VAE > RF

Ref.	Theme	Year	Application	Method	Metric
[37]	T6	2026	Conversational WQ data expert	GPT, embeddings, BlueGAP IS	Geocoding; multilingual Q&A; user feedback

The distribution of studies in the final analysis, by publication year, is as follows: 4 in 2023, 5 in 2024, 17 in 2025, and 11 in 2026. This distribution indicates a marked increase in publication activity in 2025 and 2026. The relationship between themes and years is further visualized in Figure 4.

The distributions of thematic categories, hardware/architectural components, and AI/optimization methods are jointly presented in Figure 2. Figure 2(a) shows that the studies are grouped around six main themes. In contrast, Figures 2(b) and 2(c) highlight the prominence of Raspberry Pi, ESP32, Jetson/Orin, sensor networks, edge-cloud/digital twin/blockchain architectures, CNN/vision DL, the YOLO family, classical ML, RNN/LSTM-based models, federated learning, and metaheuristic optimization approaches.

The hardware, method, and metric distributions presented in Figure 2(b) and 2(c), and in Figure 3, were calculated using a multi-labeling approach; therefore, the category totals may exceed the 37 studies included in the final analysis set.

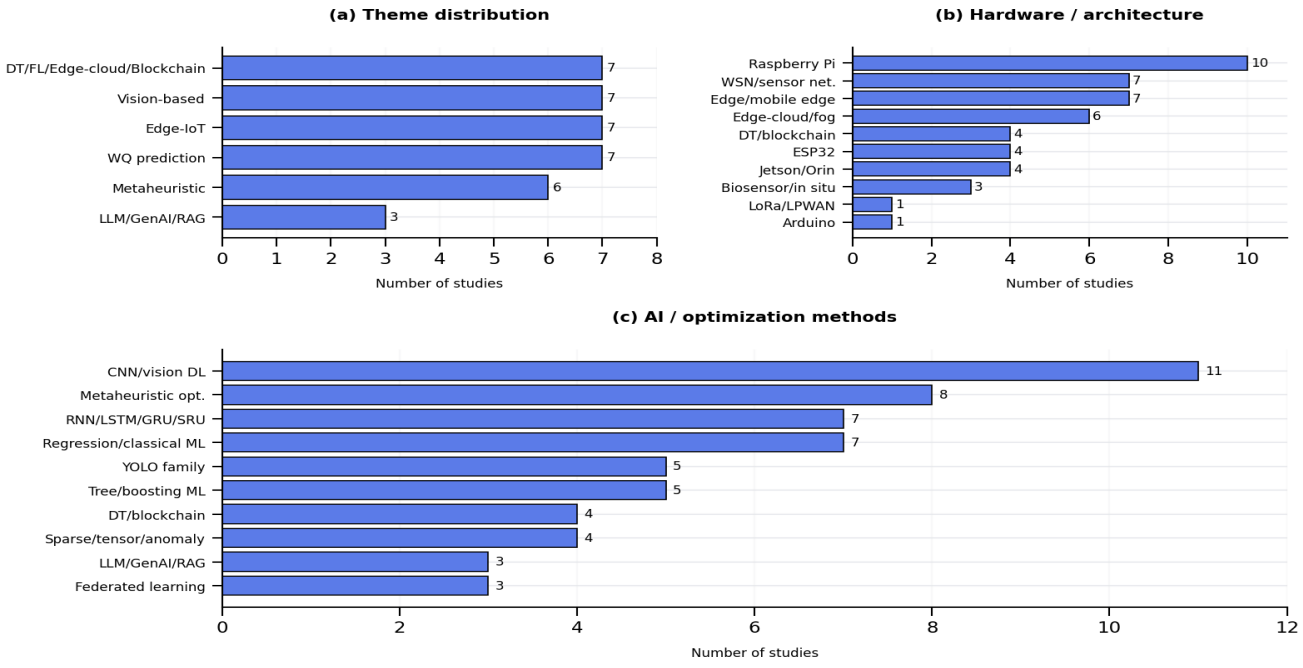


Figure 2. Combined distribution of thematic categories, hardware/architecture components, and AI/optimization methods.

Figure 3 presents the distribution of performance metrics reported in the reviewed studies. In classification-oriented studies, accuracy, precision, recall, specificity, F1-score, AUC, FPR/FNR, and misclassification cost are commonly reported. In regression, time-series, and missing-data imputation studies, RMSE, MAE, MAPE, and MSE are frequently used. The vision-based detection and segmentation studies contain mAP, AP, FPS, detection accuracy, Dice, and IoU, while distributed clustering studies include TE, QE, Silhouette Index, and

Dunn Index. Edge-IoT and system-level studies measure latency/delay, processing time, FPS, CPU/memory/power consumption, model size, GFLOPs, battery life, bandwidth/data transfer reduction, and compression ratio. Biosensor and sensor-noise studies focus on SNR, recovery, RSD, detection limit, response time, and laboratory validation discrepancy. Gas usage, transaction costs, throughput, confirmation latency, trust scores, and fault tolerance are reported in blockchain-based studies. In LLM/GenAI/RAG-based decision-support studies, metrics or evidence types such as BERTScore, FactCC, ROUGE-L, QA accuracy, Precision@1/Recall@5, decision error, and user/expert feedback are prominent [2], [4–6], [8], [10-26], [28–30], [32–37]. This diversity indicates the continuing need for common benchmarks and standardized metric sets in the literature.

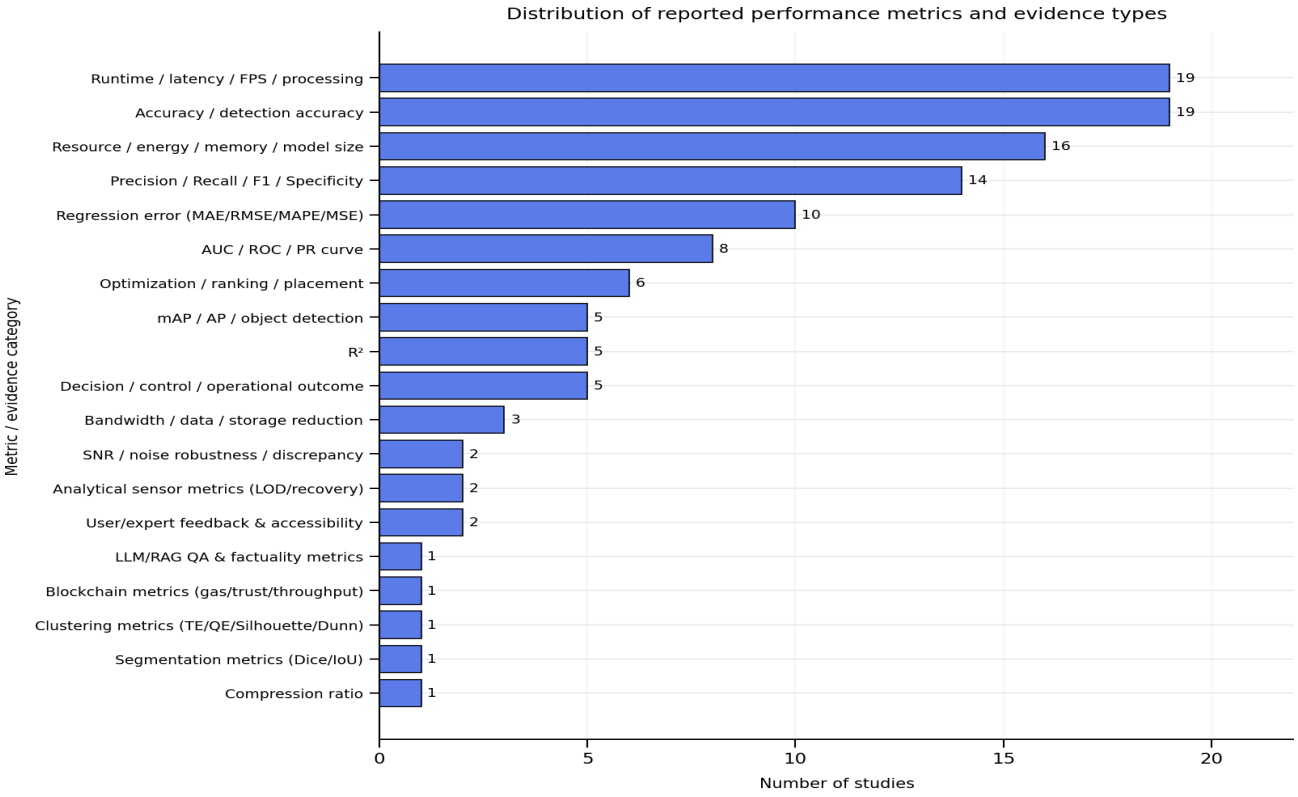


Figure 3. Distribution of reported performance metrics and evidence types.

Figure 4 shows the theme-year density map. The heatmap indicates that water quality prediction, digital twin/federated learning/edge-cloud/blockchain, and metaheuristic approaches became more prominent in 2025. In contrast, vision-based detection, LLM/GenAI/RAG decision support, and Edge-IoT studies became more visible in 2026.

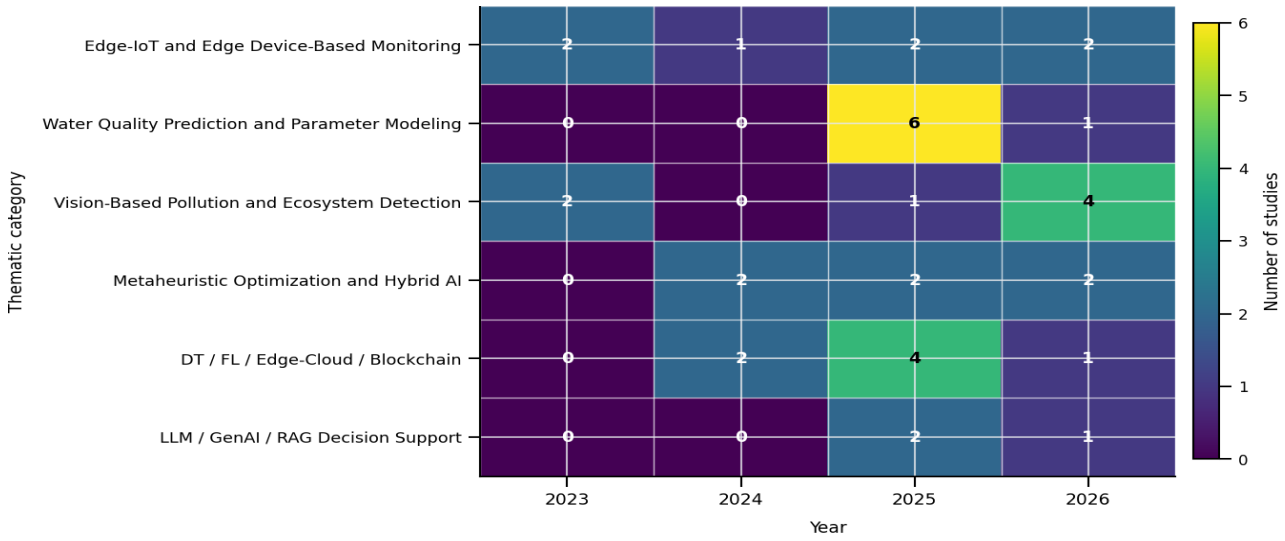


Figure 4. Theme-year density map.

V. DISCUSSION

RQ1: Edge-IoT and AIoT Architectures

Regarding RQ1, the results show that current water quality-based monitoring research is transitioning from centralized cloud-based data collection systems to local computing and decision-making systems in the field. Raspberry Pi and Jetson Orin devices are widely used in applications that require vision-based detection. Embedded deep learning is needed for autonomous control tasks [4], [5], [10], [11], [16], [21-26], while ESP32 and Arduino are commonly used in sensor-based monitoring or autonomous control tasks that require low cost. Edge-cloud, federated learning, digital twin, and distributed IoT architectures are preferred for large-scale data processing, visualization, and decision-support processes [12-14], [32-35]. This trend shows that system-level criteria such as low latency, communication overhead, local processing capability, and field deployability have become visible alongside model accuracy in the reviewed literature.

1. RQ2: AI, Optimization, and LLM-Based Methods

Regarding RQ2, the reviewed studies do not show a dominant method family; rather, the choice of method depends on the problem type, data type, hardware constraints, and real-time processing requirements. In vision-based studies [9-11], [21], [23-25], the most popular architectures are CNN/YOLO/VGG/UNet, while XGBoost, Random Forest, CatBoost, LSTM/BiLSTM, and SRU are popular architectures for sensor- and time-series-based studies [3], [6-8], [12],[13], [17-20], [28]. PSO, GA, WOA, GWO, SSO, and KBMOPSO are often used in optimization [14], [26-31], [35]. Distributed architectural approaches are featured prominently in federated learning/digital twin, and edge-cloud studies [12-14], [32-35], and natural language and decision support approaches are prominent in LLM/GenAI/RAG studies [15], [36], [37]. In addition to widely used ML/DL models, more specialized method families such as BASM-based sparse coding, ABL tensor factorization, DiSOM-based distributed clustering, VAE-based GenAI, and RAG-LLM/DQN are also notable [15], [20], [29],[30], [36],[37]. This indicates that recent studies are not limited to prediction and classification problems;

instead, they span a broader problem space that includes missing-data imputation, denoising, distributed learning, natural language-based decision support, and user interaction.

2. RQ3: Performance Metrics

Regarding RQ3, the methods of performance evaluation seem to be directly influenced by the kind of problems tackled. Widely used classification measures include accuracy, precision, recall, specificity, F1-score, and AUC. In contrast, traditional error-based measures for regression, time-series, and parameter estimation studies are R2, RMSE, MAE, MAPE, and MSE. In vision-based detection and segmentation research, metrics such as mAP, AP, FPS, detection accuracy, Dice, and IoU can be used to assess model performance and applicability in real time. For Edge-IoT and system-level studies, factors such as latency/delay, processing time, CPU/memory/power consumption, model size, GFLOPs, battery life, bandwidth reduction, and compression ratio are considered alongside algorithmic accuracy. In biosensor and sensor-noise studies, analytical reliability metrics such as SNR, recovery, detection limit, response time, RSD, and laboratory validation discrepancy are emphasized. In blockchain and secure monitoring studies, system reliability metrics such as gas usage, transaction cost, throughput, confirmation latency, trust score, fault tolerance, and false alarm reduction are used; while in LLM/GenAI/RAG-based decision-support studies, metrics or evidence types such as BERTScore, FactCC, ROUGE-L, QA accuracy, Precision@1/Recall@5, decision error, expert intervention rate, and user/expert feedback are prominent. This diversity shows that performance evaluation in the reviewed studies has expanded beyond model accuracy alone toward system-level dimensions such as real-time operation, communication overhead, resource usage, field deployability, explainability, and decision-support quality.

This also indicates that, in Edge-IoT-based water quality systems, not only model performance but also the balance among system-level criteria such as latency, energy consumption, and data reliability plays a decisive role in the design process.

3. RQ4: Research Gaps

Based on the reviewed study set, five main research gaps were identified with respect to RQ4:

- Although field/prototype validations are reported in some studies, long-term, multi-season, multi-site validations covering different environmental conditions remain limited [4], [19], [21], [25], [32],[33], [36],[37].
- Hardware resource consumption metrics that directly affect system performance, such as processing time, CPU usage, and memory usage, are not reported within a standardized and comparable framework across studies [2], [9–11], [15–18], [20], [22],[23], [28], [33], [35],[36].
- The lack of standardized datasets, open data, data quality documentation, metadata standardization, and comparable benchmarks limits direct comparison across studies [8], [14], [16], [21], [29],[30], [37].
- Studies that integrate Edge-IoT, metaheuristic optimization, federated learning/digital twin, blockchain, and LLM/GenAI/RAG approaches within the same system architecture remain limited [13], [15], [31],[32], [35–37].
- Explainability, reliability, data privacy, hallucination risk, model trustworthiness, and human-machine interaction require more systematic evaluation [12], [14],[15], [35],[37].

These gaps indicate that a single model or hardware platform cannot explain performance in water quality monitoring systems. Performance varies depending on the monitored parameter, data type, hardware capacity, energy budget, connectivity conditions, and decision-support requirements. Sensor noise, missing data, and local

adaptation of models to changing environmental conditions are important challenges for low-cost edge devices; high data volume is a key issue in vision-based tasks; and semantic interpretation and reliable decision support are emerging as important research directions for LLM/GenAI/RAG-based systems. Since this study was limited to Article-type publications indexed in the WoS Core Collection under SCI-EXPANDED and ESCI, some applied Edge-IoT prototypes published in Scopus, IEEE Xplore, the ACM Digital Library, or conference proceedings may have been outside its scope. While this choice enabled a more controlled review pool with respect to indexing quality and reproducibility, it also constitutes a natural limitation in representing the full range of engineering-oriented implementations.

VI. CONCLUSION

This study systematically reviewed Edge-IoT-based hybrid artificial intelligence studies on water quality monitoring and water pollution detection published between 2023 and 2026 in the Web of Science Core Collection under the SCI-EXPANDED and ESCI indices. Through three complementary search strategies, duplicate removal, title/abstract screening, and a five-point eligibility assessment, 37 studies [1], [37] were included in the final analysis set. The findings show that low-cost edge devices, sensor networks, vision-based deep learning models, metaheuristic optimization approaches, federated learning/digital twin/blockchain architectures, sensor denoising methods, and LLM/GenAI/RAG-based decision support systems are strongly represented in the reviewed literature on water quality monitoring. The reviewed studies indicate that no single model, method, or hardware platform is superior across all scenarios. Performance depends on the monitored parameter, data type, field conditions, hardware constraints, energy requirements, connectivity infrastructure, and decision-support needs. Therefore, future studies should consider not only model accuracy but also system-level criteria such as energy consumption, latency, bandwidth, compression ratio, device cost, edge-device resource usage, explainability, reliability, privacy, and long-term field validation. In addition, integrating LLM/GenAI/RAG-based decision-support components with edge devices, federated learning, digital twins, and blockchain-based trusted data management infrastructures may contribute to the development of next-generation hybrid water quality monitoring systems.

REFERENCES

[1]. S. O. Ajakwe, I. U. Ajakwe, T. Jun, D.-S. Kim, and J.-M. Lee, “CIS-WQMS: Connected intelligence smart water quality monitoring scheme,” Internet of Things, vol. 23, Art. no. 100800, Oct. 2023, doi: 10.1016/j.iot.2023.100800.

[2]. J. Roostaei, Y. Z. Wager, W. Shi, T. Dittrich, C. Miller, and K. Gopalakrishnan, “IoT-based edge computing (IoTEC) for improved environmental monitoring,” Sustainable Computing-Informatics & Systems, vol. 38, Art. no. 100870, Apr. 2023, doi: 10.1016/j.suscom.2023.100870.

[3]. W.-C. Hu, L.-B. Chen, B.-H. Wang, G.-W. Li, and X.-R. Huang, “Design and implementation of a full-time artificial intelligence of things-based water quality inspection and prediction system for intelligent aquaculture,” IEEE Sensors Journal, vol. 24, no. 3, pp. 3811–3821, Feb. 2024, doi: 10.1109/JSEN.2023.3340295.

[4]. P. Nuangpirom, S. Pitjamit, V. Jaikampan, C. Peerakam, W. Nakkiew, and P. Jewpanya, “Machine learning on low-cost edge devices for real-time water quality prediction in tilapia aquaculture,” Sensors, vol. 25, no. 19, Art. no. 6159, Oct. 2025, doi: 10.3390/s25196159.

[5]. Y. Rajesh, M. Dwarampudi, and Y. Pratap, “Artificial intelligence-based techniques to monitor and maintain the drinking water quality supplied to households,” International Journal of Environment and Waste Management, vol. 38, no. 3, 2025, doi: 10.1504/IJEWM.2025.149530.

[6]. R. G. Babu, T. S. Geetha, K. R. Reddy, and K. K. Kumar, "Machine learning for river water quality monitoring: assessing seasonal and agricultural influences," Environmental Earth Sciences, vol. 84, no. 21, Art. no. 626, Oct. 2025, doi: 10.1007/s12665-025-12579-5.

[7]. G. Gandhimathi, C. Chellaswamy, T. S. Geetha, and S. A. Arunmozhi, "Integrating regression and boosting techniques for enhanced river water quality monitoring in the Cauvery Basin: a seasonal and sustainable approach," Water Environment Research, vol. 97, no. 7, Art. no. e70128, Jun. 2025, doi: 10.1002/wer. 70128.

[8]. J. Gao, B. Chen, and S.-K. Tang, "Water quality monitoring: a water quality dataset from an on-site study in Macao," Applied Sciences-Basel, vol. 15, no. 8, Art. no. 4130, Apr. 2025, doi: 10.3390/app15084130.

[9]. W. Hu, Y. Zhou, J. Lu, J. Sheng, and Z. Jin, "Remote sensing detection and resource utilization of urban sewage sludge based on mobile edge computing," Ecological Chemistry and Engineering S, vol. 30, no. 2, pp. 275–282, Jun. 2023, doi: 10.2478/eces-2023-0029.

[10]. S. Li, S. Liu, Z. Cai, Y. Liu, G. Chen, and G. Tu, "TC-YOLOv5: rapid detection of floating debris on Raspberry Pi 4B," Journal of Real-Time Image Processing, vol. 20, no. 2, Art. no. 17, Apr. 2023, doi: 10.1007/s11554-023-01265-z.

[11]. H. Wu, L. Feng, X. Wang, L. Yang, and Y. Wu, "Embedded device-based oil spill measurement system for inland waters," IEEE Transactions on Instrumentation and Measurement, vol. 74, Art. no. 5046212, 2025, doi: 10.1109/TIM.2025.3613907.

[12]. C. Rodríguez-Alonso, I. Pena-Regueiro, and O. García, "Digital twin platform for water treatment plants using microservices architecture," Sensors, vol. 24, no. 5, Art. no. 1568, Mar. 2024, doi: 10.3390/s24051568.

[13]. M. K. Benhaoua, A. Dahane, R. Benameur, I. B. Benziane, H. E. Zahaf, A. Benhariga, A. Elouissi, C. Boudia, and S. Souihi, "SW3T-IoT: smart water monitoring based on a three-tier IoT architecture," IEEE Access, vol. 13, pp. 206331–206354, 2025, doi: 10.1109/ACCESS.2025.3640507.

[14]. Z. Ma, Y. Zhu, C. Chen, T. Li, Y. Li, X. Li, Y. Wang, T. D. Waite, and J. Guan, "Towards the digitalization of water treatment facilities: a case study on machine learning-enabled digital twins," Journal of Water Process Engineering, vol. 77, Art. no. 108316, Sep. 2025, doi: 10.1016/j.jwpe.2025.108316.

[15]. X. Wang, Z. Zhao, S. Rothfarb, S. Lu, L. Temple, Y. Yang, and B. Li, "Deciphering sensor reading noises in complex wastewater environments using real-time in situ hybrid GenAI wastewater-sensor entity," ACS ES&T Water, vol. 5, no. 10, pp. 5893–5906, Oct. 2025, doi: 10.1021/acsestwater. 5c00510.

[16]. S. Sharma, D. Mishra, A. Yadav, B. Gami, and E. S. Madhan, "An intelligent, low-cost water quality monitoring system with on-device machine learning and cloud integration," Scientific Reports, vol. 16, no. 1, Art. no. 11106, Feb. 2026, doi: 10.1038/s41598-026-37287-3.

[17]. G. M. Dhote, S. R. Kandavalli, K. P. Kumar, M. Madanan, G. J. Victor, and E. Muniyandy, "Intelligent carbon nanomaterial electrodes for rapid heavy metal detection in complex water matrices," Microchemical Journal, vol. 218, Art. no. 115225, Nov. 2025, doi: 10.1016/j.microc.2025.115225.

[18]. T. Jena, S. Shinde, S. E. Roselin, C. Sujana, V. S. Rao, and K. I. A. James, "Deep learning-optimized carbon quantum dot biosensors for emerging contaminant monitoring," Microchemical Journal, vol. 218, Art. no. 115389, Nov. 2025, doi: 10.1016/j.microc.2025.115389.

[19]. Supria, A. Julianto, Wahyat, Marzuarman, M. N. Faizi, and Hardiyanto, "Machine learning-based dissolved oxygen classification using low-cost IoT sensors for smart aquaponic," International Journal of Advanced Computer Science and Applications, vol. 16, no. 12, pp. 598–605, 2025.

[20]. Z. Mu, D. Li, J. Zhao, H. Jiang, F. Wu, X. Niu, Q. Wang, and S. Xue, "Boundary-aware time-frequency sparse coding for low-overhead IoT time series analytics," IEEE Wireless Communications Letters, vol. 15, pp. 2015–2019, 2026, doi: 10.1109/LWC.2026.3666680.

[21]. M. Roman, A. Zahid, A. M. Nafchi, and M. Sher, "Development of an AI-enabled vision-based on-site rapid detection of nitrates and nitrites from drainage samples," Water Air and Soil Pollution, vol. 237, no. 8, Art. no. 469, Jan. 2026, doi: 10.1007/s11270-026-09143-7.

[22]. M. Moreno, I. Trejo-Zúñiga, V. A. González-Huitrón, R. F. Santana-Cruz, R. G. García, and G. P. Chacón, "Edge node deployment for turbidity estimation in farm ponds," Big Data and Cognitive Computing, vol. 10, no. 4, Art. no. 126, Apr. 2026, doi: 10.3390/bdcc10040126.

[23]. P. Chumchu, K. Patil, and A. Nyandoro, "Real-time pond water assessment via embedded deep learning and visual data acquisition: a practical monitoring approach for aquaculture," JoVE-Journal of Visualized Experiments, no. 228, Art. no. 69744, Feb. 2026, doi: 10.3791/69744.

[24]. M. I. M. Rahman, A. F. A. Razak, A. P. P. A. Majeed, R. M. Musa, A. A. A. Jalil, I. M. Khairuddin, M. A. Abdullah, and M. A. M. Razman, "Analyzing fish detection and classification in IoT-based aquatic ecosystems through deep learning," PeerJ Computer Science, vol. 12, Art. no. e3496, Jan. 2026, doi: 10.7717/peerj-cs.3496.

[25]. D. Saranyaraj, N. Govardanan, S. N. Kumar, K. Elangovan, and R. Shashidhar, "Water hyacinth detection for autonomous navigation mapping using image segmentation cascaded classifier," Scientific Reports, vol. 16, no. 1, Art. no. 4051, Jan. 2026, doi: 10.1038/s41598-025-34128-7.

[26]. A. M. A. Bamini, R. Chitra, S. Agarwal, H. Kim, P. Stephan, and T. Stephan, "IoT-based automatic water quality monitoring system with optimized neural network," KSII Transactions on Internet and Information Systems, vol. 18, no. 1, pp. 46–63, Jan. 2024, doi: 10.3837/tiis. 2024.01.004.

[27]. A. K. Salem and A. A. Abokifa, "Sensor placement optimization in sewer networks: machine learning-based source identification approach," Journal of Water Resources Planning and Management, vol. 150, no. 11, Art. no. 4024048, Nov. 2024, doi: 10.1061/JWRMD5.WRENG-6430.

[28]. H. Hou, J. Yin, X. Qiu, J. Wu, Y. Yang, D. W. S. Tang, Y. Wang, Z. Wei, and C. Xu, "Water heavy metal prediction using a CBM model based on hybrid metaheuristic algorithms," IEEE Transactions on Geoscience and Remote Sensing, vol. 63, Art. no. 4211231, 2025, doi: 10.1109/TGRS.2025.3602639.

[29]. X. Wu, L. Wang, M. Ge, J. Jiang, Y. Cai, and B. Yang, "Adaptive biases-incorporated latent factorization of tensors for predicting missing data in water quality monitoring networks," Frontiers in Physics, vol. 13, Art. no. 1587012, Aug. 2025, doi: 10.3389/fphy. 2025.1587012.

[30]. A. Panwar and S. J. Nanda, "Distributed clustering with interpretable self-organizing maps in wireless sensor network for environmental data analysis," Environmental and Ecological Statistics, Early Access, Mar. 2026, doi: 10.1007/s10651-026-00714-9.

[31]. D. Vijay and S. Krishnan S, "Optimization and simulation of intelligent aquaponics system using evolutionary algorithms," International Journal of Limnology, vol. 62, Art. no. 4, Feb. 2026, doi: 10.1051/limn/2025013.

[32]. W. K. Cheng, J. C. Khor, W. Z. Liew, K. T. Bea, and Y.-L. Chen, "Integration of federated learning and edge-cloud platform for precision aquaculture," IEEE Access, vol. 12, pp. 124974–124989, 2024, doi: 10.1109/ACCESS.2024.3454057.

[33]. W. K. Cheng, T. B. Tan, J. S. Tan, J. Y.-L. Chan, K. T. Bea, and Y.-L. Chen, "Smart aquaculture: integrating computer vision and IoT for automated crayfish health assessment via federated learning," IEEE Access, vol. 13, pp. 189979–189994, 2025, doi: 10.1109/ACCESS.2025.3624140.

[34]. B. Das, A. Adel, T. Jan, and M. D. Wahiduzzaman, "Water quality management using federated deep learning in a developing Southeast Asian country," Water Resources Management, vol. 39, no. 4, pp. 1893–1909, Mar. 2025, doi: 10.1007/s11269-024-04051-z.

[35]. Q. Abbas, R. Almakki, M. Albathan, W. J. Obidallah, and Z. Alshaikh, "Secure water quality monitoring using blockchain-smart contract and IoT in 6G smart cities," Water Resources Management, vol. 40, no. 7, Art. no. 288, May 2026, doi: 10.1007/s11269-026-04630-2.

[36]. P. Danvirutai, S. Charoenwattanasak, S. Tola, K. Thaiso, B. Yuangsoi, H. T. Minh, and C. Srichan, "An integrating RAG-LLM and deep Q-network framework for intelligent fish control systems," Scientific Reports, vol. 15, no. 1, Art. no. 21377, Jul. 2025, doi: 10.1038/s41598-025-05892-3.

[37]. G. Vald, Y. Sermet, J. Mount, S. Shrestha, D. J. Samuel, D. Cwiertny, and I. Demir, "A novel intelligent expert system for exploring water quality data," Water Science and Technology, vol. 93, no. 4, pp. 411–427, Feb. 2026, doi: 10.2166/wst. 2026.199.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Practical Evaluation of CNN Architectures for Histopathological Lung Cancer Classification

Nazlı Çıtak
Graduate School of Natural and Applied Sciences
Gazi University
Ankara, Türkiye
nazlicitak2016@gmail.com

Ömer Faruk Bay
Faculty of Technology
Gazi University
Ankara, Türkiye
omerbay@gazi.edu.tr
ORCID:0000-0002-6823-145X

Abstract
Lung cancer remains a leading cause of cancer-related mortality worldwide, making accurate diagnostic systems critically important. Recent advances in deep learning have significantly improved histopathological image classification. However, many existing studies primarily emphasize predictive accuracy while giving less attention to training duration and practical comparison across architectures. This study presents a practical evaluation of three convolutional neural network (CNN) architectures—MobileNetV3, EfficientNet-B3, and ConvNeXt-Tiny—for histopathological lung cancer classification using the LC25000 dataset. All models were trained under a controlled, simplified experimental setup, using a fixed number of 10 epochs, a single train/validation/test split, and identical training conditions without adaptive strategies such as early stopping. Experimental results indicate that while all models achieve comparable classification performance, clear differences emerge in training duration. MobileNetV3 demonstrates the fastest training process and the most favorable accuracy–time trade-off, whereas ConvNeXt-Tiny provides a balanced performance. EfficientNet-B3 achieves the highest accuracy but requires substantially longer training time. These findings emphasize that, when classification performance is comparable, training duration becomes a decisive criterion for practical model selection.

Keywords: deep learning, CNN, histopathology, lung cancer classification

I. INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, highlighting the critical importance of early and accurate diagnosis. Histopathological image analysis plays a central role in confirming cancer types and guiding clinical decision-making. However, manual examination of histopathological slides is time-consuming and subject to interobserver variability, motivating the development of automated, reliable computational approaches. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have significantly improved the performance of medical image classification systems. CNN-based approaches have demonstrated remarkable success in histopathological image analysis by enabling automated feature extraction and high classification accuracy across various cancer types. As a result, deep learning models are increasingly considered as supportive tools in computational pathology workflows. Despite these advances, a large body of existing work primarily focuses on improving predictive performance, often under highly optimized experimental conditions. While such approaches provide valuable insights into model capability, they may not fully reflect practical experimental settings, particularly when simplified and consistent training conditions are required for fair comparison. In many cases, the feasibility of training different architectures under consistent, simplified conditions is as important as their predictive accuracy. Motivated by this gap, this study presents a practical evaluation of three widely used CNN architectures, MobileNetV3, EfficientNet-B3, and ConvNeXt-Tiny, for histopathological lung cancer classification. The models are assessed under a controlled and simplified experimental setup designed to ensure a fair comparison, using a fixed number of training epochs, a single data split, and identical training



conditions. The main objective of this study is to analyze how different CNN architectures perform in terms of both classification accuracy and training time in a simplified, controlled experimental setting. By focusing on a simplified, controlled setup, the study aims to provide clear, interpretable insights into the trade-off between predictive performance and training duration.

II. RELATED WORK

Recent advances in deep learning have significantly transformed medical image analysis, particularly in cancer diagnosis. Convolutional neural networks (CNNs) have emerged as the dominant paradigm due to their ability to automatically learn hierarchical feature representations from large-scale data, outperforming traditional machine learning methods across a wide range of applications [1], [2]. Architectural innovations such as depth scaling, multi-path structures, and attention mechanisms have further enhanced CNN performance. In lung cancer diagnosis, deep learning models have demonstrated strong performance across radiological and histopathological imaging modalities. EfficientNet-based approaches, for instance, provide a balanced scaling strategy that improves both accuracy and efficiency [3], [4]. Raza et al. proposed Lung-EffNet, achieving high classification accuracy while maintaining relatively efficient training characteristics.

Recent studies on lightweight CNNs have shown that reduced-complexity architectures can maintain strong classification performance [5]. Similarly, XLLC-Net demonstrated that compact architectures with fewer parameters can maintain high accuracy while significantly reducing training time and improving interpretability [6]. Hybrid and multi-model approaches have also been explored to enhance classification performance. Naveenraj and Vijayakumar (2025) proposed a hybrid DCNN–LSTM framework achieving high accuracy, albeit with increased computational demands [7]. Attallah (2025) introduced a multi-CNN feature integration approach that combines multiple architectures and feature selection techniques to improve performance while reducing complexity [8]. In addition, data augmentation techniques have been shown to improve model robustness and classification accuracy [9]. Modern CNN designs such as ConvNeXt further demonstrate that convolution-based architectures remain competitive with transformer-based models while maintaining efficiency and scalability [10]. Despite these advances, most studies primarily emphasize classification accuracy, often reporting near-saturated performance on benchmark datasets. However, relatively limited attention has been given to training duration and practical efficiency considerations.

In contrast, this study focuses on a controlled comparison of representative CNN architectures under identical conditions, emphasizing training duration and accuracy–time trade-off as key criteria for model selection in practical scenarios.

III. MATERIAL AND METHODS

1. Dataset

This study utilizes the publicly available LC25000 dataset, which contains histopathological images of lung and colon tissues. In this work, only the lung subset is considered, including three classes: lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. All images are resized to a consistent input resolution and normalized to ensure uniformity across models. The dataset is divided into training, validation, and test sets using a fixed split of 80%, 10%, and 10%, respectively. This single split configuration is adopted to maintain a controlled and consistent evaluation setting.

2. Model Architectures

In this study, three representative convolutional neural network architectures, MobileNetV3, EfficientNet-B3, and ConvNeXt-Tiny, were selected to reflect different design philosophies and performance characteristics in modern deep learning. MobileNetV3 was chosen as a lightweight and fast architecture designed for rapid training and early convergence. EfficientNet-B3 is a more complex scaling-based architecture that aims to achieve high classification performance by increasing model capacity. ConvNeXt-Tiny, a modern architecture inspired by transformer-based design principles, offers a balanced approach by combining contemporary architectural improvements with competitive performance. By selecting these models, the study aims to provide a practical comparison across lightweight, high-capacity, and modern balanced architectures.

3. Training Configuration

All experiments were conducted on a computing system equipped with an Intel Core i7-13700H CPU. All models are trained under a simplified, controlled experimental setup, with a fixed number of 10 epochs for each architecture. Identical training conditions are applied across all models, including preprocessing steps, input dimensions, and optimization settings. This controlled configuration enables a direct comparison of architectural performance without introducing variability from training strategies.

4. Evaluation Metrics

Model performance is evaluated using both predictive and training-related metrics. Classification performance is measured using accuracy on the test set.

$$\text{Accuracy-Time Trade-off} = \text{Test Accuracy} / \text{Total Training Time}$$

Higher values indicate more favorable predictive performance relative to training duration. Trade-offs between predictive performance and training efficiency are commonly analyzed in deep learning studies. To explicitly quantify this relationship, a simple accuracy–time trade-off metric is defined as a compact measure of training efficiency. Thus, models with higher values achieve better predictive efficiency per unit of training time. In addition, total training time is recorded for each model. By jointly considering accuracy and training duration, this study provides a practical perspective on model comparison, emphasizing the balance between predictive performance and training duration. This evaluation framework enables a more interpretable comparison of models by capturing both performance and training efficiency.

IV. RESULTS AND DISCUSSION

The results show that all models achieve high classification performance, with test accuracy values exceeding 0.99. EfficientNet-B3 yields the highest accuracy (0.9987), followed by ConvNeXt-Tiny (0.9967) and MobileNetV3 (0.9913). However, substantial differences are observed in training duration (Table 1) and are illustrated in Figure 1.

Table 1. Performance comparison of CNN architectures on the LC25000 dataset.

Model	**Test Accuracy**	**Avg Epoch Time (s)**	**Total Time (s)**	**Accuracy -Time Trade-off**

MobileNetV3	0.9913	450.81	4508.10	0.000220
EfficientNet-B3	0.9987	1387.83	13878.30	0.000072
ConvNeXt-Tiny	0.9967	1122.02	11220.24	0.000089

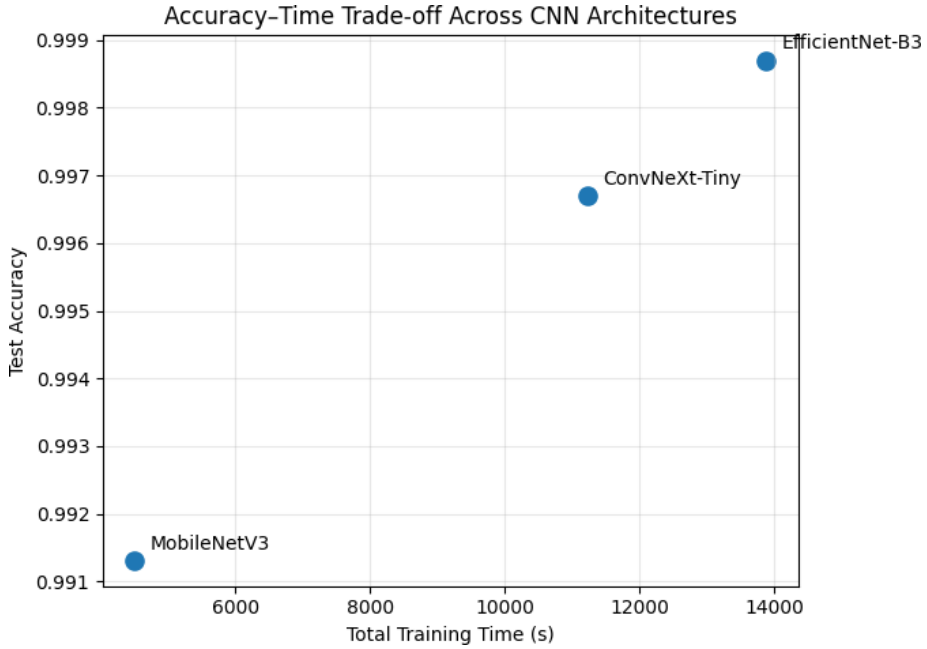
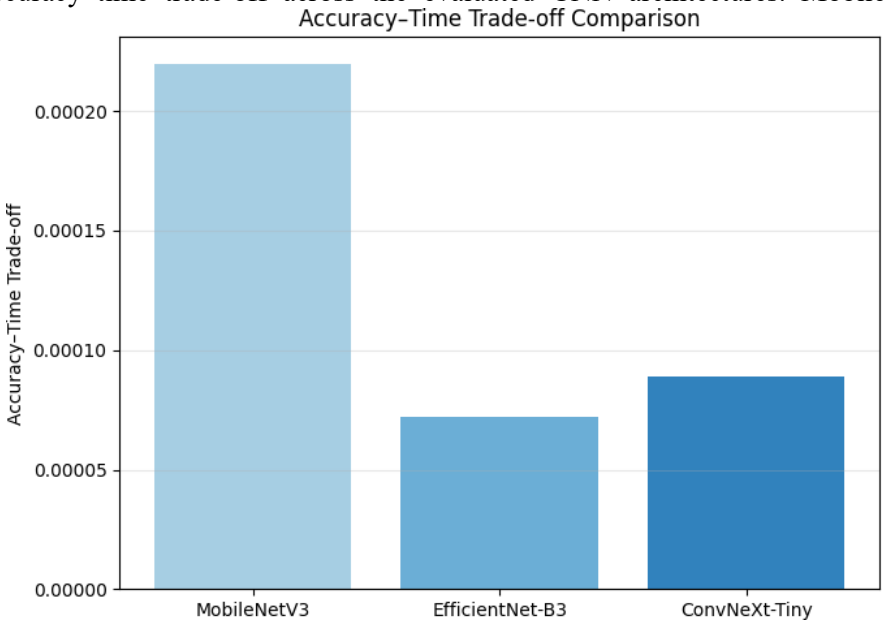


Figure 1. Accuracy–time trade-off across the evaluated CNN architectures. MobileNetV3 achieves competitive accuracy with significantly reduced training time, whereas EfficientNet-B3 provides only marginal accuracy improvements at substantially higher training time. ConvNeXt-Tiny demonstrates a balanced performance between these two extremes.

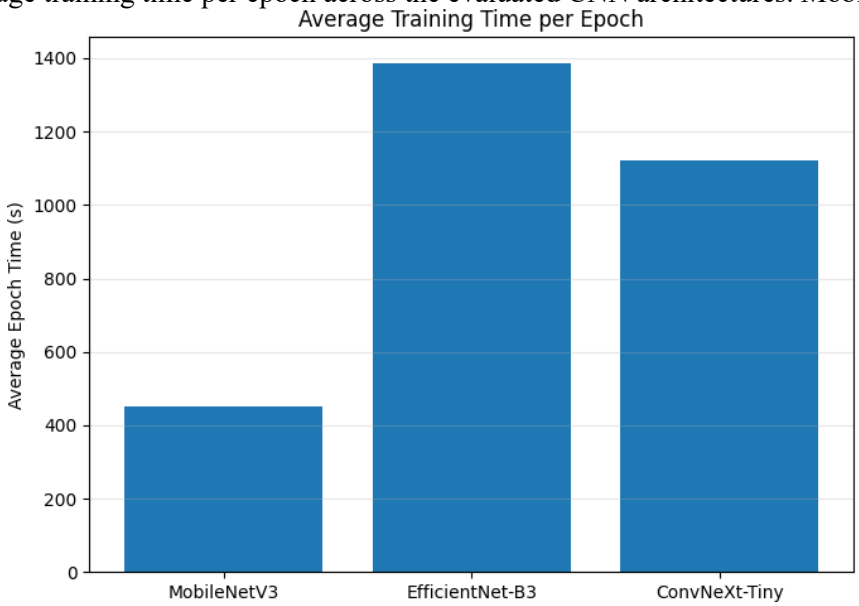
The accuracy–time trade-off metric further highlights these differences, as clearly presented in Figure 2. MobileNetV3 achieves the highest trade-off value (0.000220), indicating the most efficient accuracy–time balance. In contrast, EfficientNet-B3 yields the lowest trade-off value due to its substantially longer training time, despite achieving the highest accuracy. ConvNeXt-Tiny provides a moderate trade-off, reflecting its balance between accuracy and training duration. Overall, the results show that differences in predictive performance are relatively small across models. In contrast, differences in training duration are substantial, indicating that model selection should not rely solely on accuracy, particularly under practical constraints.

Figure 2. Accuracy–time trade-off across the evaluated CNN architectures. MobileNetV3 achieves the highest trade-off value, indicating the most efficient accuracy–time balance.



MobileNetV3 achieves the shortest training time, with an average epoch duration of 450.81 seconds, significantly outperforming the other models in training efficiency. In contrast, EfficientNet-B3 requires more than three times as long per epoch (1387.83 seconds), while ConvNeXt-Tiny shows an intermediate training duration (1122.02 seconds), as shown in Figure 3.

Figure 3. Average training time per epoch across the evaluated CNN architectures. MobileNetV3 achieves significantly shorter training time compared to the other models.



MobileNetV3 emerges as the most efficient model, offering competitive accuracy while significantly

reducing training time. Its stable validation behavior across epochs further supports its suitability for practical deployment scenarios.

EfficientNet-B3 achieves the highest accuracy; however, this improvement is marginal compared with the significant increase in training duration. This suggests that the additional training time may not be justified in scenarios where training efficiency is critical.

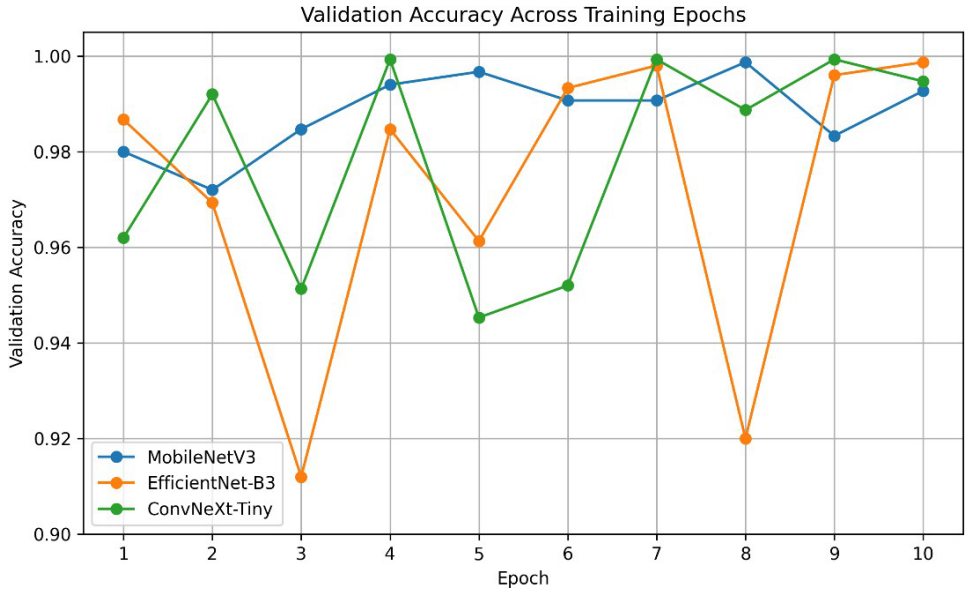


Figure 4. Validation accuracy across training epochs for the evaluated CNN architectures. MobileNetV3 demonstrates relatively stable validation behavior, whereas EfficientNet-B3 and ConvNeXt-Tiny exhibit greater variability across epochs. MobileNetV3 demonstrates relatively stable validation behavior throughout training, whereas EfficientNet-B3 and ConvNeXt-Tiny exhibit greater fluctuations across epochs, as illustrated in Figure 4.

ConvNeXt-Tiny achieves high peak validation accuracy but exhibits noticeable variability across epochs, indicating less consistent generalization than MobileNetV3. While it provides a balance between accuracy and training duration, its fluctuating validation performance suggests reduced stability under the given experimental setup. Overall, the findings emphasize that training duration and performance consistency are critical factors in model selection, particularly when accuracy differences are marginal.

V. CONCLUSION

This study demonstrates that while all evaluated models achieve high classification accuracy, their training characteristics differ substantially. MobileNetV3 provides the best balance between accuracy and training time, making it the most efficient choice in the evaluated setting. In contrast, EfficientNet-B3 achieves the highest accuracy; however, this gain is marginal compared with the significant increase in training time. ConvNeXt-Tiny demonstrates strong performance but exhibits less consistent validation behavior across epochs. Overall, the results indicate that maximizing accuracy alone is not sufficient for effective model

selection. Instead, when predictive performance is comparable, training duration becomes a decisive factor in practice, favoring models that achieve competitive accuracy with shorter training time and stable performance.

REFERENCES

[1]. L. Alzubaidi, M. Al-Amidie, J. Zhang, et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," Journal of Big Data, vol. 8, no. 1, p. 53, 2021.

[2]. A. Khan, A. Sohail, U. Zahoora, et al., "A survey of the recent architectures of deep convolutional neural networks," Artificial Intelligence Review, vol. 53, pp. 5455–5516, 2020.

[3]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.

[4]. R. Raza, F. Zulfiqar, M. O. Khan, et al., "Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images," Engineering Applications of Artificial Intelligence, vol. 126, part B, p. 106902, 2023.

[5]. A. Raza, F. Hanif, and H. A. Mohammed, "Clinical validation of lightweight CNN architectures for reliable multi-class classification of lung cancer using histopathological imaging techniques," Scientific Reports, vol. 16, no. 1, p. 6512, 2026.

[6]. J. R. Jim, M. E. Rayed, M. F. Mridha, and K. Nur, "XLLC-Net: A lightweight and explainable CNN for accurate lung cancer classification using histopathological images," PLOS ONE, vol. 20, no. 5, e0322488, 2025.

[7]. M. Naveenraj and P. Vijayakumar, "Enriched lung cancer classification approach using an optimized hybrid deep learning approach," Scientific Reports, vol. 15, p. 37573, 2025.

[8]. O. Attallah, "Lung and colon cancer classification using multiscale deep features integration of compact convolutional neural networks and feature selection," Technologies, vol. 13, no. 2, p. 54, 2025.

[9]. V. Karunakaran, N. Akshaya, B. Harismita, and S. Atchaya, "Lung cancer classification using CNN with data augmentation," in Proc. ICICNIS, 2024, pp. 1152–1157.

[10]. Z. Liu, H. Mao, C.-Y. Wu, et al., "A ConvNet for the 2020s," in Proc. CVPR, 2022, pp. 11966–11976.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Edge-Based Digital Twin for Real-Time Energy Monitoring in Food Manufacturing: A Case Study of a Bakery Enterprise

Bauyrzhan Amirkhanov
Faculty of Information Technologies and Artificial Intelligence
Al-Farabi Kazakh National University
Almaty, Kazakhstan
amirkhanov.b@gmail.com
ORCID:0000-0002-4915-0347

Murat Kunelbayev
Institute of Information and Computational Technologies CS MES RK
Almaty, Kazakhstan
murat7508@yandex.kz
ORCID:0000-0002-5648-4476

Miras Tokhtassyn
Faculty of Information Technologies and Artificial Intelligence
Al-Farabi Kazakh National University
Almaty, Kazakhstan
mirastoktasyn30@gmail.com
ORCID:0009-0003-8682-8262

Alina Raeva
Faculty of Information Technologies and Artificial Intelligence
Al-Farabi Kazakh National University
Almaty, Kazakhstan
alinaraevali18@gmail.com
ORCID:0009-0001-4257-6919

Gulshat Amirkhanova
Faculty of Information Technologies and Artificial Intelligence
Al-Farabi Kazakh National University
Almaty, Kazakhstan
gulshat.aa@gmail.com
ORCID:0000-0003-3933-5476

Abstract
Energy overconsumption in food manufacturing represents a significant operational burden, particularly in resource-constrained regions. This paper presents DigitalEgiz, an edge-based digital twin system designed for real-time energy monitoring in a commercial bakery enterprise in Almaty, Kazakhstan. The system integrates a distributed network of 18 ESP32 microcontrollers, two SAIMAN smart electricity meters (Dala and Orman models), and a Raspberry Pi 5 edge gateway, which communicates via MQTT to a cloud-hosted stack comprising InfluxDB, Grafana, and OpenTwins. Over a 61-day deployment period, the system collected 17.7 million data points with an adjusted data completeness of 94.19%. Analysis revealed that 81% of monitored devices operate below Kazakhstan's mandatory 0.93 power factor threshold, exposing facilities to regulatory fines. The system also demonstrated that predictive models (a Prophet+LSTM hybrid achieving MAE = 3.39 kWh) integrated within the digital twin can provide real-time decision support. Results confirm that edge-based digital twins offer a cost-effective and scalable path toward energy optimization and regulatory compliance in industrial IoT settings.

Keywords: digital twin; energy monitoring; IoT; edge computing; food manufacturing; ESP32; InfluxDB; Grafana; power factor; Kazakhstan

I. INTRODUCTION

The global food manufacturing sector accounts for a substantial share of industrial energy consumption. Yet, a disproportionate fraction of this energy is wasted due to unmonitored equipment behavior, poor power factor, and reactive rather than proactive maintenance strategies [1]. In emerging economies such as Kazakhstan, where energy tariffs are regulated and infractions carry administrative penalties, the need for intelligent energy management systems is acute. Digital twins—virtual replicas of physical systems that synchronize with real-world sensor data in real time—have emerged as a transformative paradigm for industrial monitoring and optimization [2]. When combined with edge computing principles, they enable low-latency data processing closer to the source, reducing bandwidth requirements and improving system responsiveness [3]. While digital twin deployments in energy-intensive industries such as petrochemicals and automotive manufacturing have been extensively studied, their application in small- to medium-sized food processing enterprises, particularly in Central Asia, remains largely unexplored. This paper addresses this gap by presenting DigitalEgiz, a purpose-built edge-based digital twin deployed at a commercial bakery enterprise (referred to as GoldSapa) in Almaty, Kazakhstan. The primary contributions of this work are: (1) a heterogeneous IoT architecture combining microcontroller-based distributed sensing with industrial-grade smart meters; (2) a lightweight MQTT-to-cloud pipeline hosted on a Raspberry Pi 5 edge gateway; (3) a 61-day empirical evaluation capturing 17.7 million data points; and (4) integration of a Prophet+LSTM hybrid forecasting model within the digital twin interface for decision support.

II. RELATED WORK

The concept of digital twins in manufacturing was first formalized by Grieves [4] and has since been applied across numerous industrial domains. In energy management, Zhou et al. [5] demonstrated that digital twins reduce energy waste in smart factories by enabling continuous monitoring and anomaly detection. Similarly, Wen et al. [6] showed that integrating IoT sensing with digital twin platforms improves predictive maintenance accuracy in food processing lines. Edge computing has become a critical enabler for IoT-based industrial systems, as evidenced by Shi et al. [3], who proposed an edge-cloud collaborative framework for real-time industrial analytics. The combination of edge processing with time-series databases such as InfluxDB has been shown to support high-throughput sensor ingestion at sub-second latency [7]. In the domain of energy forecasting, hybrid models that combine classical statistical approaches (e.g., SARIMA, Prophet) with deep learning (e.g., LSTM) have demonstrated superior accuracy compared to single-model approaches [8]. Prophet, developed by Facebook (now Meta), is particularly well-suited to energy time series with strong seasonal patterns and irregular holidays, while LSTM networks capture non-linear temporal dependencies [9]. Despite this rich literature, no prior work has addressed the specific challenges of deploying digital twins in bakery enterprises within Kazakhstan's regulatory framework, where power factor compliance (PF $\geq$ 0.93) carries statutory financial penalties. This work fills this gap.

III. SYSTEM ARCHITECTURE

1. Hardware Layer

The sensing layer of DigitalEgiz consists of 18 ESP32 microcontrollers (ESP-WROOM-32 modules) deployed across production equipment, including industrial ovens, proofing chambers, mixing machines, and HVAC units. Each ESP32 samples voltage (RMS), current (RMS), active power, reactive power, apparent power, and power factor at 1-second intervals using ACS712 current sensors and ZMPT101B voltage transformers. Two SAIMAN smart electricity meters are integrated at the facility level: the SAIMAN Dala model (single-phase) and the SAIMAN Orman model (three-phase). These meters provide authoritative aggregate consumption data

via the RS-485 Modbus protocol, which the Raspberry Pi 5 gateway reads over a USB-to-RS-485 adapter. All 18 ESP32 nodes communicate wirelessly over the facility's 2.4 GHz Wi-Fi network.

2. Edge Gateway and Communication

The Raspberry Pi 5 (8 GB RAM) serves as the central edge gateway. It hosts a Mosquitto MQTT broker that aggregates telemetry from all ESP32 nodes, subscribing to a structured topic hierarchy of the form facility/zone/device/metric. A Python-based MQTT-to-InfluxDB bridge written using the paho-mqtt and influxdb-client libraries forwards incoming messages to the cloud-hosted InfluxDB instance with a median end-to-end latency of 47 ms. The gateway also executes local anomaly detection logic: when a device reports a power factor below 0.93 for more than 60 consecutive seconds, a real-time alert is generated and forwarded to the Grafana alerting engine. This local processing reduces cloud API calls by approximately 40% compared to a purely cloud-centric architecture.

3. Cloud Platform and Digital Twin Visualization

The cloud layer is hosted on a VPS running Ubuntu 22.04, deploying InfluxDB v2.7 as the time-series database, Grafana v10.2 as the visualization and alerting layer, and OpenTwins [10] as the digital twin management framework. OpenTwins provides a Unity 3D WebGL-based 3D model of the bakery floor plan, overlaid with real-time sensor readings and color-coded status indicators. Grafana dashboards are organized into functional panels: (1) real-time energy consumption per device; (2) power factor compliance status; (3) peak-hour identification; (4) deviation from baseline charts; and (5) a decision support panel ("Decision Support") that displays forecasting outputs and actionable recommendations. The InfluxDB data retention policy is set to 90 days for raw data and indefinite for hourly aggregates. Table 1 summarizes the key specifications of the deployed system.

Table 1. Key System Specifications

Parameter	Value
Sensor nodes	18 ESP32 + 2 SAIMAN
Sampling rate	1 second
Deployment period	61 days
Total data points	17.7 million
Data completeness	94.19% (adjusted)
Edge gateway	Raspberry Pi 5 (8 GB)
MQTT broker	Mosquitto v2.0
Time-series DB	InfluxDB v2.7
Visualization	Grafana v10.2
Digital twin engine	OpenTwins (Unity WebGL)
E2E latency (median)	47 ms

IV. ENERGY FORECASTING MODULE

The forecasting component of DigitalEgiz employs a hybrid Prophet+LSTM architecture to generate 24-hour-ahead energy consumption predictions for each monitored device. Prophet models the trend and seasonality components (daily, weekly, and holiday effects relevant to the Kazakhstani production calendar). At the same time, the LSTM network (two stacked layers of 64 units each, with dropout = 0.2) captures residual non-linear

dynamics. Training data consisted of 45 days of hourly aggregate consumption per device zone. The Prophet model was fitted first, and its residuals were fed as input sequences to the LSTM. Final predictions were obtained by summing the Prophet forecast with the LSTM-predicted residuals. The model was evaluated on the remaining 16 days using Mean Absolute Error (MAE) as the primary metric. The Prophet+LSTM hybrid achieved MAE = 3.39 kWh on the held-out test set, outperforming standalone Prophet (MAE = 4.82 kWh) and standalone LSTM (MAE = 4.11 kWh). A SARIMA+LSTM baseline achieved MAE = 3.71 kWh, confirming the advantage of the Prophet-based decomposition for this production schedule pattern. Forecasting outputs are published to InfluxDB every 15 minutes and rendered in the Grafana "Decision Support" panel, which displays predicted vs. actual consumption curves. Anomaly flags are triggered when deviations exceed 15% of the predicted value. Operators can use these flags to schedule equipment maintenance or load-shifting actions.

V. EXPERIMENTAL RESULTS

4. Data Collection and Quality

The system was deployed continuously from day 1 to day 61. Raw data completeness (including facility-wide power outages) stood at 87.3%. After excluding nine documented outage events totaling approximately 52 hours—during which all devices were powered off simultaneously, so no partial data loss occurred—the adjusted data completeness reached 94.19%. This metric more accurately reflects the reliability of the sensing infrastructure, independent of external power-supply conditions. A duplicate device entry (Baker1_001989 and ESP32_Dala_Meter_001989 referring to the same physical unit) was identified and resolved during post-processing, correcting aggregate consumption totals by approximately 2.1%. All results reported in this paper are based on the corrected dataset.

Figure 1 shows the distribution of hourly energy consumption recorded over the full deployment period.

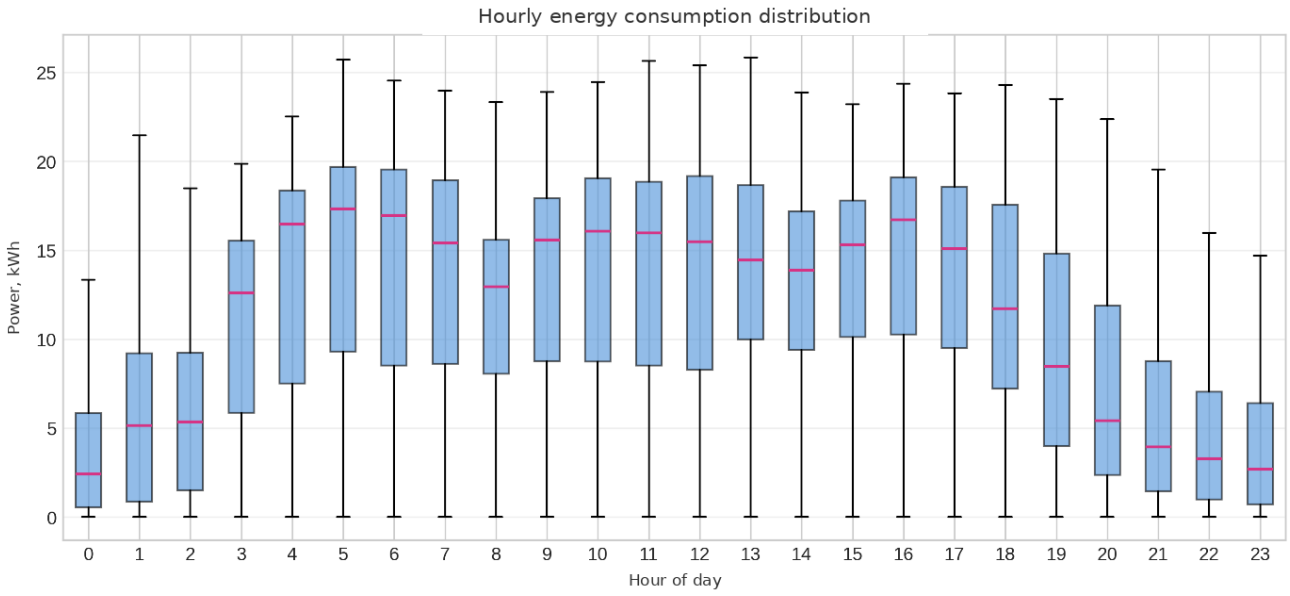


Figure 1. Hourly energy consumption distribution (kWh) across the 61-day deployment period.

5. Power Factor Analysis

Power factor (PF) is a critical compliance indicator under Kazakhstan's energy regulatory framework [11]. Facilities with PF below 0.93 are subject to fines of approximately 1,638 USD for the first violation (calculated based on the Monthly Calculation Index, MCI, in effect at the time of deployment). Analysis of the 61-day dataset revealed that 81% of monitored devices (14.6 of 18 devices on average) operated below the 0.93 threshold during at least one monitoring window per day. The most severe non-compliance was observed in the industrial oven zone (average PF = 0.71) and the mixing machine bank (average PF = 0.76). The HVAC units showed the highest compliance rate, operating above 0.93 in 89% of measured intervals. These findings underscore the practical utility of continuous PF monitoring via the digital twin dashboard.

6. System Performance

The edge gateway maintained 99.4% uptime over the 61-day deployment period, with three brief restarts due to OS-level package updates. The average CPU utilization of the Raspberry Pi 5 during normal operation was 23%, with peak utilization of 61% during batch ingestion events. Memory utilization averaged 31% (2.5 GB of 8 GB), confirming that the hardware is well-dimensioned for the workload. The InfluxDB instance stored approximately 4.2 GB of raw time-series data, growing at roughly 70 MB/day.

VI. DISCUSSION

The results confirm that an edge-based digital twin can be effectively deployed in a real-world food manufacturing setting with minimal infrastructure investment. The Raspberry Pi 5 gateway proved adequate for handling the aggregate telemetry from 18 sensor nodes and two industrial meters at 1-second resolution, while simultaneously running MQTT brokering, anomaly-detection logic, and forecasting inference. The 81% power-factor non-compliance rate is a striking finding with direct financial implications for the enterprise. While the system itself cannot correct power factor without additional power conditioning hardware (such as capacitor banks), the digital twin provides the visibility needed to identify which equipment zones require intervention. This directly supports a data-driven investment decision: installing a 50 kVAR capacitor bank for the oven zone alone would bring that zone into compliance and avoid the regulatory fine. The hybrid Prophet+LSTM forecasting model demonstrated clear advantages over individual models, validating the decomposition approach for bakery production schedules, which exhibit strong weekly patterns tied to product-type rotations and irregular holiday shutdowns. The 15% deviation threshold for anomaly alerting was calibrated empirically and proved sensitive enough to detect equipment degradation events while maintaining an acceptable false-positive rate. One limitation of the current deployment is its reliance on a Wi-Fi network for ESP32 communication, which can cause occasional packet loss in areas with poor coverage. Future work will explore adding a secondary ZigBee mesh network as a fallback communication channel. Additionally, the OpenTwins 3D visualization currently requires a WebGL-capable browser, which limits mobile accessibility; a lightweight REST-based mobile interface is planned.

VII. CONCLUSION

This paper presented DigitalEgiz, an edge-based digital twin system for energy monitoring in a bakery enterprise. The system successfully integrated 18 ESP32 sensor nodes, two industrial smart meters, and a Raspberry Pi 5 edge gateway into a coherent IoT-to-cloud pipeline, collecting 17.7 million data points over 61 days with 94.19% adjusted data completeness. Key findings include widespread power factor non-compliance (81% of devices are below the 0.93 regulatory threshold) and effective energy forecasting via a Prophet+LSTM hybrid model,

achieving an MAE of 3.39 kWh. The deployment demonstrates that cost-effective, open-source digital twin architectures are feasible for small- to medium-sized food manufacturing enterprises in Central Asia. The system provides actionable insights for regulatory compliance, predictive maintenance, and operational optimization. Future work will extend the platform to multi-site deployments and incorporate reinforcement learning-based demand response strategies.

ACKNOWLEDGMENT

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992975).

REFERENCES

[1]. International Energy Agency, "Food and agriculture: Key energy trends to 2050," IEA Report, Paris, France, 2023.
[2]. M. Grieves, "Digital twin: Manufacturing excellence through virtual factory replication," White Paper, 2014.
[3]. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016.
[4]. M. Grieves and J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," in Transdisciplinary Perspectives on Complex Systems, Springer, Cham, 2017, pp. 85–113.
[5]. Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," Proc. IEEE, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.
[6]. L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on ResNet-50," Neural Comput—Appl., vol. 32, pp. 6111–6124, 2020.
[7]. InfluxData, "InfluxDB documentation: Time series platform," [Online]. Available: https://docs.influxdata.com. [Accessed: June 2025].
[8]. B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in Proc. ICLR, 2020.
[9]. S. J. Taylor and B. Letham, "Forecasting at scale," Am. Stat., vol. 72, no. 1, pp. 37–45, 2018.
[10]. OpenTwins Project, "OpenTwins: An open-source digital twin framework," [Online]. Available: https://github.com/ertis-research/opentwins. [Accessed: June 2025].
[11]. Committee for Technical Regulation and Metrology of Kazakhstan, "Rules for electricity supply and use," Order No. 202, Nur-Sultan, Kazakhstan, 2019.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Machine Learning Models for Monitoring and Analysis of Mobile Object Movements at a Food Industry Enterprise

Bauyrzhan Amirkhanov
Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
amirkhanov.b@gmail.com
ORCID:0000-0002-4915-0347

Yenlik Faruzkyzy
Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
elnura.dauletkhan@gmail.com
ORCID:0009-0001-1030-8266

Alikhan Amirkhanov
Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
alikhan.amirkhan@gmail.com
ORCID:0009-0003-3708-3153

Gulshat Amirkhanova
Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
gulshat.aa@gmail.com
ORCID:0000-0003-3933-5476

Dina Baizhanova
Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
dina.baizhanova.2024@gmail.com
ORCID:0000-0002-6109-8174

Abstract
Food safety standards require food enterprises to control the movement of personnel and mobile equipment through production areas, yet most plants still rely on periodic manual inspections. This paper describes a video analytics pipeline that converts footage from a single ceiling camera at a working food plant into zone-level movement indicators for two object classes, workers and mobile containers. The pipeline combines a YOLO11 detector fine-tuned on a small-domain dataset, kinematic multi-object tracking, and a spatial layer that maps trajectories to five production zones. Evaluation used a temporal data split with a held-out test hour. YOLO11x with 22 frozen layers reached mAP50 of 0.939 on the held-out hour; deeper fine-tuning scored higher on validation but lost up to 0.10 mAP50 on the test hour. ByteTrack matched BoT-SORT on association quality (HOTA 0.783 against 0.767) and ran substantially faster (27.6 against 17.0 frames per second). A full one-hour run produced interpretable occupancy, entry, and dwell-time statistics for each production zone. The experiments also document a feature-forgetting effect that appears when a fine-tuned model is retrained on an isolated annotation round, and show that consolidating the rounds and retraining from base weights removes it.

Keywords: computer vision, object detection, multi-object tracking, YOLO11, ByteTrack, ceiling camera, zone analytics, food industry

I. INTRODUCTION

Food safety management standards, in particular HACCP and ISO 22000:2018, require food enterprises to control the movement of personnel and material flows in production areas: to separate clean and dirty zones, to record how long workers stay at critical control points, and to document compliance with sanitary protocols [1]. At most plants, this control still depends on periodic walk-through inspections with checklists. Inspections of

this kind cover only a small part of the working day, depend on the individual auditor, and leave no continuous record of events [2], [3].

Computer vision mitigates this limitation by providing automatic, contactless, and reproducible monitoring. Industrial deployments already exist: a neural network trained on 500 images of tool wear eliminated about 30% of unplanned downtime associated with manual wear control [4], and a machine vision system integrated with industrial controllers reduced false positives by 28% while maintaining detection accuracy above 95% [5]. In the food sector, a ResNet50 model achieved 97.9% accuracy in classifying meat products [6], and a modified YOLOv8 achieved an mAP50 of 0.75 when detecting 28 categories of personal protective equipment [7]. These systems analyze the product or the protective gear. How people and equipment actually move through the shop floor has received far less attention. Ceiling cameras are a natural choice for movement monitoring in a production hall. The overhead view gives a near-orthographic projection of the floor and removes horizontal occlusion between objects. The same view, however, degrades the appearance features that detectors learn from standard datasets: models trained on horizontal views lose up to 55.7% of relative mAP when applied to overhead imagery [8]. Kinematic trackers that rely on motion models rather than appearance continue to work under these conditions [9], [10]. Still, there is little published evidence of their behavior on real footage from food plants.

This paper presents a processing pipeline that converts video from a ceiling camera at a working food plant into zone-level movement indicators for two object classes: workers and mobile production containers. The contributions are the following. First, the paper compares YOLO11 model scales and fine-tuning depths on real overhead footage, with all decisions evaluated on a temporally held-out test hour rather than on validation frames adjacent to the training data. Second, it documents a feature-forgetting effect that appears when an already-adapted detector is fine-tuned again on an isolated annotation round, and shows that consolidating the rounds and retraining from base weights removes the effect. Third, it compares the ByteTrack and BoT-SORT trackers under identical detections on a static overhead scene. Fourth, it describes a zone analytics module that converts trajectories into occupancy, entry, exit, and dwell-time indicators, and validates the module on a full one-hour run.

II. RELATED WORK

Most computer vision research in the food industry targets the product. Reviews of machine vision in food production cover quality assessment, contaminant detection, and HACCP-oriented monitoring of animal-source foods [2], [3]. Compliance applications add detection of protective equipment [7]. Movement of staff and equipment within the production area appears only marginally in these reviews, although food safety audits specifically require such records. Single-stage detectors in the YOLO family dominate real-time detection applications. Surveys trace the architecture from the first version to YOLO11 [11], [12]; the YOLO11 generation introduces the C3k2 block and the C2PSA attention module and ships in five scales from nano to extra large [13]. For multi-object tracking, the tracking-by-detection paradigm remains standard [14]. ByteTrack associates low-confidence detections in a second matching stage and recovers partly occluded objects without any appearance model [9]. BoT-SORT extends this scheme with camera motion compensation and re-identification features [10]. The HOTA metric balances detection and association quality, complementing MOTA and IDF1 [15]. Overhead and fisheye monitoring is a separate research line. Surveys of person detection from the overhead view [16], [17] and specialized methods such as RAPiD [18] show that the top-down perspective changes object geometry enough to break models trained on horizontal data; the WEPDTOF dataset documents this for in-the-wild ceiling cameras [19]. Appearance-based re-identification is weak in this setting [20]. A recent warehouse study combined ceiling cameras with a digital twin for asset tracking [21], which is the setting closest to ours, although it does not address food production constraints. Adapting a pretrained detector to a small domain dataset

raises two practical questions: how many layers to freeze and how to cheaply grow the dataset. Layer-freezing studies for YOLO architectures report that partial freezing often beats full fine-tuning on small datasets [22], [23], and active learning reduces manual annotation effort by an order of magnitude in some domains [24]. On the analytics side, trajectory-based behavior recognition has been applied in retail [25] and industrial scenarios [26], usually with heatmaps and zone statistics as the output layer. A pipeline that integrates overhead detection, kinematic tracking, and zone analytics for a food plant has not been described before; this is the gap this work addresses.

III. MATERIALS AND METHODS

1. Pipeline architecture

The system is a sequence of connected layers: video input, detection, multi-object tracking, spatial semantics, indicator extraction, and analytics output (Figure 1). The detector localizes objects of the two classes; the tracker links detections into trajectories with persistent identities; the spatial layer assigns each trajectory to production zones; and the final layer computes occupancy, entries, exits, and dwell times. The detector and the tracker are replaceable components. The work does not claim that the selected models are universally optimal; it shows that, for this overhead scene and this experimental protocol, the selected combination yields the most stable balance between quality and speed. The reusable part is the selection and verification scheme on temporally held-out data.

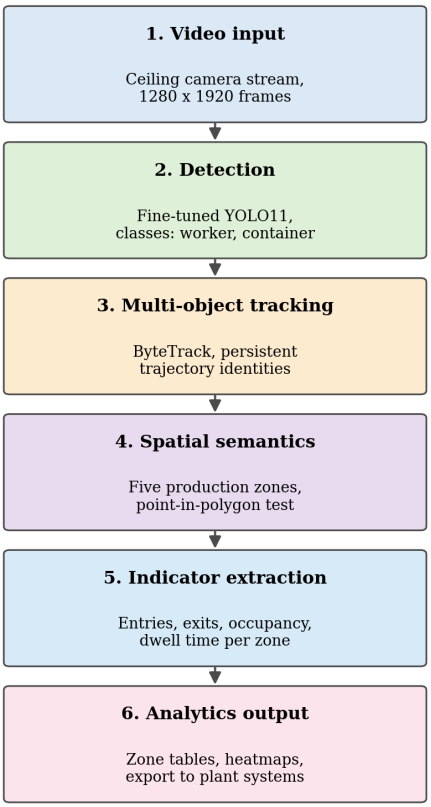


Figure 1. Overall scheme of the movement monitoring pipeline: from the video stream to analytical movement indicators.

2. Production scene and data protocol

The experiments use recordings from one ceiling camera installed above the production area of a food plant. The camera recorded four one-hour files at a frame resolution of 1280 by 1920 pixels during a single working day. The scene contains workers, mobile containers, racks, central equipment, and waiting and loading areas. Two production-relevant classes are annotated: a worker, which reflects operational activity, and a container, which reflects the movement and accumulation of production objects. In the annotation tooling, the classes were labeled person and vat. The main experiment applies a temporal data split. The first two hours form the training pool, the third hour serves for validation, and the fourth hour remains an independent held-out test and the source of all tracking and zone analytics runs. The split addresses a specific leakage risk: with a random frame split, neighboring near-duplicate frames from the same hour end up in both the training and test sets, inflating the scores. A model evaluated this way looks better than it will behave in the next hour of real footage.

3. Dataset and annotation

Annotation was semi-automatic: a pretrained model proposed draft boxes, and an operator corrected them in an interactive interface, thereby keeping manual effort low without sacrificing label quality [24]. The process ran as an active learning loop. The model trained on a small set of corrected frames produced better drafts for the next batch, and frames with systematic errors were added to the training set as hard negatives. Table I lists the resulting sets. For tracking evaluation, two 30-second fragments, each containing 180 frames, were manually annotated with persistent identities.

Table 1. Composition of the experimental datasets

Set	Frames	Objects	Source
Training	120	433	hours 1 and 2
Validation	60	231	hour 3
Main test	150	593	held-out hour 4
Hard cases	60	231	held-out hour 4
Tracking fragment 1	180	1054	hour 4, 7 identities
Tracking fragment 2	180	899	hour 4, 5 identities

4. Detector training protocol

All five YOLO11 scales, from nano to extra large, start from weights pretrained on MS COCO and train under identical conditions: input size 1280 pixels, SGD with momentum 0.937 and weight decay 5e-4, initial learning rate 0.001 with cosine decay, 5 warm-up epochs, 50 epochs in total, and 22 frozen layers as the base mode. The large input size matters for the overhead view: at lower resolutions, a worker occupies less than 32 by 32 pixels and falls into the small-object category. Augmentation is conservative to limit overfitting on the small dataset: mosaic with a 0.5 ratio that switches off 10 epochs before the end of training, moderate HSV color jitter, affine transforms limited to a 10-degree rotation and 0.3 scale, and horizontal flips. For the largest model, two further fine-tuning modes are tested: 10 frozen layers and full unfreezing. The selection criterion is quality on the held-

out test hour, not the best validation score, because in operation, the model encounters new time intervals rather than frames adjacent to those in its training data.

5. Zone model and metrics

In every frame, the detector outputs a set of boxes with class labels and confidence scores, and the tracker converts them into trajectory states with identities that persist across frames. Persistent identity is what makes individual dwell times and correct transition events possible. Each zone z is a polygon P_z in frame coordinates. For a trajectory k with reference point $p_k(t)$, the box center in the current implementation, zone membership on frame t is the indicator function

$$m_{k,z}(t) = I[p_k(t) \in P_z] \quad (1)$$

The dwell time of object k in zone z equals the sum of the indicator over frames divided by the video frame rate. An entry event fires when the object was outside the zone on the previous frame and inside on the current one; an exit is the reverse transition. Detection quality is measured with precision, recall, mAP50, and mAP50-95; tracking quality with HOTA, MOTA, IDF1, the number of identity switches, and processing speed [14], [15].

6. Tracking and zone analytics protocol

ByteTrack and BoT-SORT run on identical detections from the best fine-tuned model: confidence threshold 0.25, IoU threshold 0.5, input size 1280. The lowered confidence threshold lets the second association stage of ByteTrack recover partly occluded objects from low-confidence detections [9]. The camera is static, so camera motion compensation in BoT-SORT has nothing to correct, and appearance-based re-identification is weak in the overhead view [16], [20]. Zone analytics uses five production zones: a left transit zone, a rack buffer zone, an equipment service zone, a front container waiting zone, and a right loading zone. Presence heatmaps accumulate box centers and smooth them with a Gaussian kernel. The implementation is in Python with PyTorch, Ultralytics, OpenCV, NumPy, and SciPy; training and inference were run on an NVIDIA RTX 4090 GPU with 24 GB of memory.

IV. RESULTS AND DISCUSSION

1. Domain gap and feature forgetting

The stock COCO model fails on this scene, which quantifies the domain gap. YOLO11x with original COCO weights detected workers with a maximum confidence of 0.48, below the standard 0.5 cut-off, and labeled the production containers as umbrella and toilet, the COCO classes whose top-down projections resemble a cylindrical vat. Fine-tuning is therefore not optional for the overhead view; without it, the detector misses the very objects the monitoring task is built around.

A preliminary prototype, built on a separate recording from the same plant, ran the annotation loop in three rounds and exposed an effect that matters for any iterative labeling workflow. The first model, trained on 150 frames, reached mAP50 of 0.968. Fine-tuning this already-adapted model on an isolated second round of 100 new frames dropped mAP50 to 0.910 and worker recall from 0.852 to 0.764, indicating the model lost part of what it had learned in the first round. Consolidating both rounds into a single set of 250 frames and retraining from the base COCO weights restored mAP50 to 0.972 and increased worker recall to 0.897 (Table 2, Figure

2). The practical rule is direct: do not fine-tune sequentially on isolated annotation rounds; merge the rounds and retrain from base weights [22], [23].

Table 2. Prototype metrics across active learning rounds

Metric	v1 (150 frames)	v2 (isolated 100)	v3 (merged 250)
mAP50	0.968	0.910	0.972
mAP50-95	0.804	0.750	0.797
Worker P / R	0.919 / 0.852	1.000 / 0.764	0.960 / 0.897
Container P / R	1.000 / 0.957	0.930 / 1.000	0.957 / 0.981

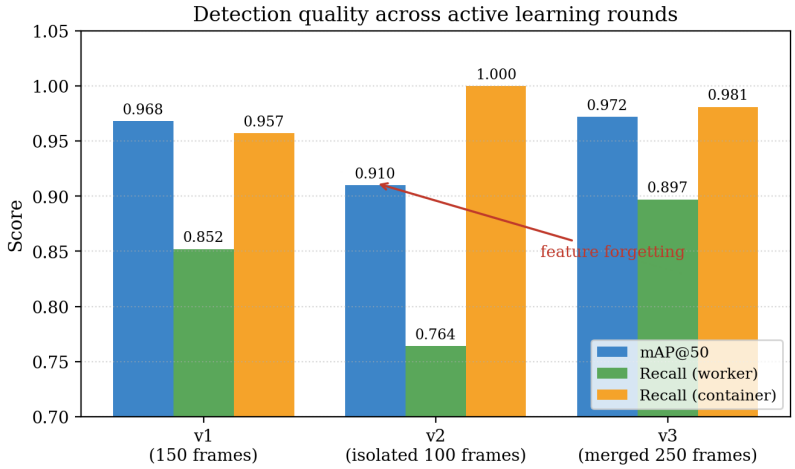


Figure 2. Metric dynamics across the active learning rounds v1, v2, and v3 of the prototype.

2. Comparison of YOLO11 scales

Table III compares the five model scales on the held-out main test. YOLO11x leads with mAP50 of 0.939 and mAP50-95 of 0.754 at 18.2 ms per frame. YOLO11l follows at 0.928 and 0.645, and YOLO11s remains the strongest compact option at 0.901 and 0.603 with a 4.3 ms inference time. The hard-case set confirms the ranking: YOLO11x reaches mAP50 of 0.936 and mAP50-95 of 0.821 there, while YOLO11l and YOLO11m tie at 0.927. The results support a three-tier deployment recommendation: YOLO11x when quality takes priority, YOLO11l as a lighter model with similar quality, and YOLO11s under hard compute constraints. The container class is recognized with almost no errors (0.97 in the normalized confusion matrix); the worker class limits overall quality due to motion and partial overlap near the equipment.

Table 3. Detector comparison on the held-out main test

Model	P	R	mAP50	mAP50-95	ms/frame
YOLO11n	0.933	0.780	0.869	0.559	2.2

YOLO11s	0.935	0.830	0.901	0.603	4.3
YOLO11m	0.908	0.797	0.900	0.598	9.4
YOLO11l	0.918	0.871	0.928	0.645	11.9
YOLO11x	0.883	0.884	0.939	0.754	18.2

3. Effect of fine-tuning depth

The fine-tuning depth experiment gave the most instructive result of the study: the best validation metric does not guarantee the best behavior on a new hour. The mode with 10 frozen layers won on validation with an mAP50 of 0.954, above the base mode's 0.927. On the held-out main test, the same model dropped to 0.894, and full unfreezing fell further to 0.836 (Table IV). The small dataset and the specific background allow a deeply adapted model to fit local features quickly, which raises the validation score but worsens transferability to a new time interval. Freezing 22 layers preserves most of the pretrained representation and adapts only part of the model; it was the most stable mode on both held-out sets. The methodological lesson for industrial adaptation: a model must not be selected on validation data that is temporally adjacent to its training data; a separate temporal test that simulates future operation is needed.

Table 4. YOLO11x fine-tuning modes

Mode	**Set**	**mAP50**	**mAP50-95**
freeze 22	validation	0.927	0.591
freeze 22	main test	0.939	0.754
freeze 22	hard cases	0.936	0.821
freeze 10	validation	0.954	0.600
freeze 10	main test	0.894	0.603
freeze 10	hard cases	0.906	0.628
full unfreeze	validation	0.909	0.495
full unfreeze	main test	0.836	0.523
full unfreeze	hard cases	0.791	0.540

4. ByteTrack against BoT-SORT

For a static overhead camera, the heavier association machinery of BoT-SORT brings no gain. Across the two annotated fragments, ByteTrack achieved HOTA of 0.783, MOTA of 0.718, and IDF1 of 0.787, with 15 identity switches, at 27.6 frames per second. BoT-SORT scored 0.767, 0.719, and 0.786 with 17 switches at 17.0 frames per second (Table V). The MOTA difference is negligible, but ByteTrack is slightly ahead on HOTA and IDF1 and processes frames 1.6 times faster, which decides the choice for stream processing. All identity switches occurred in the first fragment, which contains interactions and partial overlaps; on the simpler second fragment, both trackers ran without a single switch. Identity errors, therefore, concentrate in dense scenes and remain the main quality reserve.

Table 5. Tracker comparison on the annotated fragments

Tracker	**HOTA**	**MOTA**	**IDF1**	**IDSW**	**FPS**
ByteTrack	0.783	0.718	0.787	15	27.6

BoT-SORT	0.767	0.719	0.786	17	17.0

5. Zone analytics on the one-hour run

In the two short annotated clips, the zone variables behaved consistently with what an observer sees in the footage: the front container waiting zone dominated both fragments, with mean occupancies of 2.34 and 3.11 objects, and the transit zone remained quiet under the box-center rule. The final check applied the selected detector and tracker to the full held-out hour with the five-zone layout (Fig. 3). The equipment service zone carried the main load: 276 entries, a mean occupancy of 1.36 objects, an occupancy share of 95.18%, and a total dwell time of 4912 seconds. The front container waiting zone came second with 204 entries and 4001 seconds of dwell time; the remaining three zones stayed peripheral (Table VI). Fig. 4 shows a frame from this run with zones and trajectories overlaid. For production analysis, this output is no longer raw surveillance video but a spatial load map of the hall over an operational period.

Table 6. Zone statistics for the full one-hour run

Zone	Entries	Exits	Mean occ.	Occup., %	Dwell, s
left transit	185	83	0.153	13.34	550
rack buffer	248	52	0.437	28.50	1572
equipment service	276	172	1.364	95.18	4912
front waiting	204	112	1.111	68.90	4001
right loading	168	56	0.286	26.18	1029

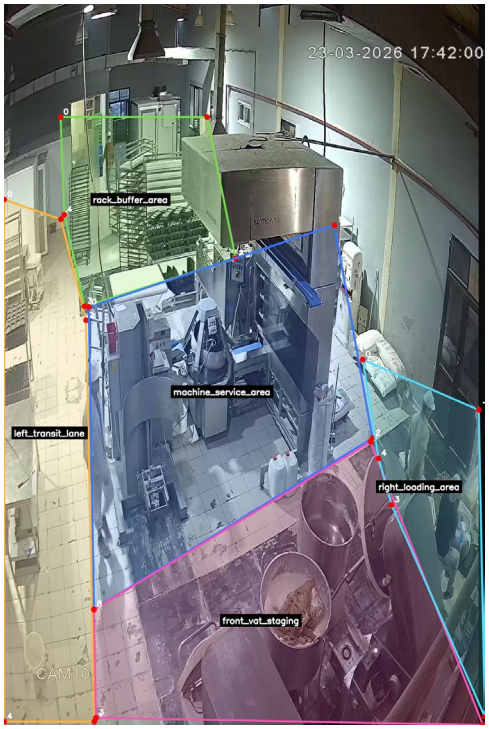


Figure 3. Layout of the five production zones for the analyzed ceiling camera.

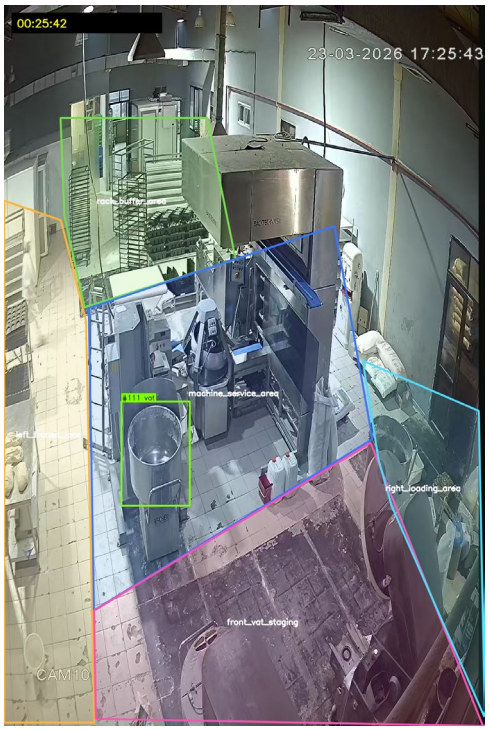


Figure 4. Frame from the full one-hour run with zones and trajectories overlaid.

6. Limitations

Four limitations bound the conclusions. The study uses a single camera at a single site; the temporal split strengthens internal validity but does not replace checks across other halls, lines, and recording days. The box-center membership rule is simple and reproducible, but for a worker, the box center does not always match the point of contact with the floor, which distorts transitions at zone borders; a bottom-center reference point for workers and the box center for containers would better fit the geometry. The hour-long run produced 639 unique track identifiers, which reflect trajectory fragmentation rather than the true number of physical objects: aggregate occupancy and dwell statistics remain valid, but per-object accounting over long horizons requires track stitching or a more robust re-identification. Finally, the pipeline prepares movement indicators but does not yet demonstrate a full integration cycle with a plant information system.

V. CONCLUSION

The paper described and verified a machine learning pipeline that monitors the movements of workers and mobile containers at a food plant from a single ceiling camera. Overhead footage becomes a set of structured movement indicators when detection, kinematic tracking, spatial semantics, and zone analytics are combined into a single reproducible chain. Three findings transfer to similar industrial settings. A stock COCO detector fails on the overhead scene, and additional fine-tuning during an isolated annotation round quietly erodes already-learned features; merging the rounds and retraining from base weights restores quality. The best validation score is a poor selection criterion under a temporal split: the mode with 22 frozen layers lost on validation but won on the held-out hour with mAP50 of 0.939. At the same time, deeper adaptation overfitted

the small dataset. For a static overhead camera, the simpler ByteTrack gives a better balance of association quality and speed than BoT-SORT.

The detection and tracking core runs faster than the input frame rate, so the pipeline suits real-time processing, and the zone indicators are ready for transfer to a plant analytics system. Planned next steps address the observed weak points: class-dependent reference points for zone membership, track stitching to reduce identity fragmentation on long horizons, additional cameras and recording days to test external validity, and a defined exchange format for integration with production information systems.

REFERENCES

[1]. A. J. Adeyemo, “Integrated HACCP and ISO 22000 frameworks for enhanced food safety compliance,” Int. J. Res. Sci. Innov., 2025.
[2]. S. Babaie and N. Norouzi, “AI-powered HACCP systems for enhanced food quality and safety assurance: a brief review,” 2024.
[3]. B. Yin, G. Tan, R. Muhammad et al., “AI-powered innovations in food safety from farm to fork,” Foods, vol. 14, no. 11, 2025.
[4]. M. Dziubek, J. Rysiński, and D. Jancarczyk, “Exploring the ViDiDetect tool for automated defect detection in manufacturing with machine vision,” Appl. Sci., vol. 13, no. 19, 2023.
[5]. M. Maślanka, D. Jancarczyk, and J. Rysiński, “Integration of machine vision and PLC-based control for scalable quality inspection in Industry 4.0,” Sensors, vol. 25, no. 20, 2025.
[6]. L. E. Flores-Zuazo, E. A. Paredes-Molina, and E. A. Carrera-Salas, “Deep learning-based quality control system for meat products,” Cogent Eng., 2025.
[7]. A. A. Mohy, H. A. Bassioni, E. O. Elgendi et al., “Deep learning enabled computer vision model for automated safety compliance in construction environments,” J. Inf. Technol. Constr., vol. 30, 2025.
[8]. A. Mao, C. Xu, O. Wang et al., “COCO-O: a benchmark for object detectors under natural distribution shifts,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 6304–6315.
[9]. Y. Zhang, P. Sun, Y. Jiang et al., “ByteTrack: multi-object tracking by associating every detection box,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022.
[10]. N. Aharon, R. Orfaig, and B. Bobrovsky, “BoT-SORT: robust associations multi-pedestrian tracking,” arXiv:2206.14651, 2022.
[11]. M. Kotthapalli, D. Ravipati, and R. Bhatia, “YOLOv1 to YOLOv11: a comprehensive survey of real-time object detection innovations and challenges,” arXiv:2508.02067, 2025.
[12]. J. Terven, D. Córdova-Esparza, and J. Romero-González, “A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS,” Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1680–1716, 2023.
[13]. R. Khanam and M. Hussain, “YOLOv11: an overview of the key architectural enhancements,” arXiv:2410.17725, 2024.
[14]. A. Milan, L. Leal-Taixé, I. Reid et al., “MOT16: a benchmark for multi-object tracking,” arXiv:1603.00831, 2016.
[15]. J. Luiten, A. Osep, P. Dendorfer et al., “HOTA: a higher order metric for evaluating multi-object tracking,” Int. J. Comput. Vis., vol. 129, pp. 548–578, 2021.
[16]. J. Konrad, M. Cokbas, M. O. Tezcan et al., “Overhead fisheye cameras for indoor monitoring: challenges and recent progress,” Front. Imaging, vol. 3, 2024.
[17]. M. Ahmad, I. Ahmed, K. Ullah et al., “Person detection from overhead view: a survey,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, 2019.

[18]. Z. Duan, M. O. Tezcan, H. Bahng et al., "RAPiD: rotation-aware people detection in overhead fisheye images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2020.

[19]. M. O. Tezcan, Z. Duan, H. Bahng et al., "WEPDTOF: a dataset and benchmark algorithms for in-the-wild people detection and tracking from overhead fisheye cameras," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2022, pp. 1123–1132.

[20]. M. Cokbas, J. Bolognino, J. Konrad et al., "FRIDA: fisheye re-identification dataset with annotations," in Proc. IEEE Int. Conf. Adv. Video Signal-Based Surveillance (AVSS), 2022.

[21]. J. Cheng, C. Verhulst, P. De Clercq et al., "Real-time warehouse monitoring with ceiling cameras and digital twin for asset tracking and scene analysis," Logistics, vol. 9, no. 4, 2025.

[22]. V. Gandhi and S. Gandhi, "Fine-tuning without forgetting: adaptation of YOLOv8 preserves COCO performance," arXiv:2505.01016, 2025.

[23]. A. D. Dobrzycki, A. M. Bernardos, and J. R. Casar, "An analysis of layer-freezing strategies for enhanced transfer learning in YOLO architectures," Mathematics, vol. 13, no. 15, 2025.

[24]. P. Bangert, H. Moon, J. O. Woo et al., "Active learning performance in labeling radiology images is 90% effective," Front—Radiol., vol. 1, 2021.

[25]. M. Shili, S. Jayasingh, and S. Hammedi, "Advanced customer behavior tracking and heatmap analysis with YOLOv5 and DeepSORT in retail environment," Electronics, vol. 13, no. 23, 2024.

[26]. H. Wang, T. Song, Z. Xu et al., "A visual trajectory-based method for personnel behavior recognition in industrial scenarios," Sensors, vol. 25, no. 20, 2025.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Explainable Mammography Classification via Attention-Guided SE-EfficientNet-B3 and Grad-CAM Diagnostic Interpretation.

Abdine Ben Ali
Institute of Informatics
Gazi, University
Ankara, Türkiye
23802002007@gazi.edu.tr

Sinan Toklu
Faculty of Technology
Gazi University
Ankara, Türkiye
stoklu@gazi.edu.tr

Abstract

Timely and accurate detection of malignant masses in mammographic images remains a critical, unsolved challenge in clinical radiology. Although advanced deep learning-based computer-aided diagnosis methods have been developed, their clinical deployment remains limited due to the "black-box" nature of high-capacity networks and limited robustness to inter-scanner variability. In this paper, we present SEE-B3, a framework that embeds Squeeze-and-Excitation (SE) channel attention directly into an EfficientNet-B3 backbone and couples it with Gradient-weighted Class Activation Mapping (Grad-CAM) to produce both high-accuracy predictions and spatially faithful diagnostic overlays. The framework is evaluated on a large-scale, multi-source breast mass corpus comprising 24,576 images drawn from the INbreast, MIAS, and DDSM collections, each independently preprocessed with CLAHE contrast enhancement before integration. SEE-B3 achieves 91.4% accuracy, an AUC-ROC of 0.946, sensitivity of 90.2%, and specificity of 92.6%, outperforming seven comparative architectures. Grad-CAM heatmaps generated by the SE-recalibrated backbone achieve a Pointing Game Accuracy of 84.7%, substantially improving lesion co-localization relative to the unattended baseline, faithfully highlighting spiculated mass margins and calcification clusters consistent with ACR BI-RADS descriptors. These results suggest that co-designing channel attention and gradient-based explainability within a single compact network yields diagnostic transparency without compromising classification performance. The proposed framework advocates clinically transparent, high-confidence AI-assisted mammography interpretation with direct relevance to standardized medical AI deployment pipelines.

Keywords: *Breast cancer, explainable AI, Grad-CAM, EfficientNet-B3, Squeeze-and-Excitation*.

I. INTRODUCTION

Breast cancer continues to be the leading cause of cancer death in women; it is estimated that 2.3 million new cases and over 685,000 deaths occurred globally in 2023 [1]. Malignant breast masses account for the highest cancer burden in women across most global regions, and the gap between early-stage detection and mortality reduction is well-documented: organized mammographic screening has been shown to lower breast cancer mortality by 20-40% over three decades of follow-up [2]. Yet the pathway from a suspicious mammogram to a confident radiological decision remains probabilistic, shaped by equipment-specific acquisition artifacts, tissue density variation, and substantial inter-reader disagreement ranging from 8% to 36% across institutions [3]. Computer-aided detection and diagnosis (CADe/CADx) in mammography has witnessed significant progress with convolutional neural networks (CNNs). The regulatory pathway for AI-based mammography systems demands both demonstrable performance and mechanical transparency. EU MDR 2017/745 and FDA SaMD guidance for AI/ML software require post-market performance monitoring and interpretability documentation that current black-box CNN systems cannot readily supply. When a network produces a confident malignancy

prediction without a spatially grounded explanation, the radiologist is placed in the position of either overriding the system, negating its benefit, or accepting it uncritically, introducing liability and patient safety concerns. [4].

Grad-CAM [5] addresses the explainability problem by producing class-discriminative spatial saliency maps; however, the quality of the heatmaps is highly dependent on the discriminative power of the learned representations. Coarse features produce diffuse overlays that are anatomically uninformative. In this paper, we propose a unified SEE-B3 framework that co-designs CLAHE-enhanced preprocessing, Squeeze-and-Excitation (SE) feature refinement, and Grad-CAM visualization to simultaneously maximize classification performance and diagnostic interpretability. The main contributions are:

The technical contributions of this work are threefold. First, CLAHE-enhanced input representations are shown to produce sharper, more anatomically coherent Grad-CAM gradients, establishing a measurable link between preprocessing fidelity and explainability quality. Second, SE recalibration is injected at stages 5-6 of EfficientNet-B3, suppressing pectoral muscle and skin-line channels in favor of lesion-correlated activations. This channel selection mechanism improves Pointing Game Accuracy by 11.6 percentage points over the unaugmented backbone. Third, a structured qualitative analysis of Grad-CAM overlays against ACR BI-RADS morphological descriptors validates that high-confidence malignant predictions are grounded in spiculated margins and calcification cluster textures. In contrast, benign predictions concentrate within smooth-bordered mass bodies, providing a degree of radiological face validity rarely demonstrated in automated mammography classification.

An SE-augmented EfficientNet-B3 backbone with channel-wise recalibration that produces sharper, more pathology-focused gradient signals under Grad-CAM.

Qualitative Grad-CAM heatmap analysis providing side-by-side comparison of baseline and refined localization patterns against radiologically expected lesion morphology.

II. RELATED WORK

1. CNN-Based Mammography Classification.

Early applications of CNN to mammography involved fine-tuning AlexNet and VGG architectures on small corpora [6]. The ResNet and DenseNet families increased representational depth and gradient flow to enable richer multi-scale feature hierarchies [7]. EfficientNet compounds scaling of depth, width, and resolution by compound coefficients, and achieves competitive results on medical imaging benchmarks [8]. Shen et al. [9] proposed a dual-pathway architecture that fuses global context and local ROI patches, achieving an AUC of 0.91 on INbreast. Nevertheless, these architectures are opaque predictors without any spatial justification.

2. Attention and Explainability in Medical AI

Hu et al. demonstrated that CNNs benefit from an explicit channel reweighting step that first compresses spatial information into a compact descriptor via pooling, and then learns a low-rank excitation function to independently modulate each channel's contribution [10]. Applied to mammography, this excitation mechanism has direct clinical value: channels encoding pectoral muscle, fatty parenchyma, or skin lines can be actively downweighted, while channels sensitive to mass-margin texture and density inhomogeneity are amplified before the classifier. CBAM [11] improved this by adding joint channel and spatial attention. Grad-CAM [5] generates class-discriminative saliency by weighting activation maps with globally averaged class gradients. Kim et al.

[12] used Grad-CAM for BI-RADS prediction, achieving 78% spatial overlap with radiologist annotations. A remaining gap is the co-design of preprocessing, attention, and explainability into a single, clinically deployable framework.

III. METHODOLOGY

1. Dataset

Experiments are conducted on the publicly available breast mammography corpus introduced by Lin and Huang [13], which consolidates mass-annotated acquisitions from three geographically and technologically distinct institutions. The constituent sources, INbreast from Hospital de São João (Porto), the MIAS collection from the UK Mammographic Society, and the CBIS-DDSM repository from the University of South Florida, vary substantially in detector generation, spatial resolution, and patient demographics, providing a heterogeneous corpus that approximates the multi-scanner variability encountered in real-world screening programs. The authors of the dataset applied CLAHE enhancement and offline augmentation independently to each source image before merging, resulting in a final combined corpus of 24,576 images at a native resolution of 227×227 pixels. For SEE-B3, images are resized bilinearly to 224×224 to conform to EfficientNet-B3's compound-scaled input requirements. All images carry binary image-level annotations: benign mass or malignant mass. Table I summarizes the per-source composition.

Table 1. Dataset of Breast Mammography Images with Masses

Source	Masses	After Aug.	Augm. Factor
INbreast	106	7,632	×72
MIAS	53	3,816	×72
DDSM	2,188	13,128	×6
Combined [13]	**2,347**	**24,576**	

2. PREPROCESSING PIPELINE

The CLAHE preprocessing is already part of the Mendeley distribution. CLAHE enhancement is already incorporated into the distributed corpus by the dataset authors; no additional contrast processing is applied during training. It is worth noting that CLAHE's contextual histogram equalization, operating on localized square tiles with a clip threshold that prevents runaway contrast amplification in texturally uniform parenchyma, is particularly well-matched to mammographic mass morphology. Spiculated borders, microcalcification clusters, and density transitions at mass peripheries are precisely the high-spatial-frequency structures that benefit from localized, clip-limited equalization, which explains the empirically observed improvement in SE attention selectivity when CLAHE-processed inputs are used. The grayscale mammograms are converted to three-channel RGB by repeating the channels to meet the ResNet50 input requirements while maintaining full intensity fidelity. Because the Mendeley corpus ships with a×6×72 offline augmentation already applied at the source level, any additional training-time augmentation must be calibrated carefully to avoid compounding artificial variability. Overly aggressive geometric distortions on an already augmented distribution can produce morphological configurations with no analog in clinical acquisition, potentially training the network on spurious texture patterns rather than true lesion morphology. SEE-B3 therefore restricts online training transforms to a minimal set: random horizontal mirroring ($p = 0.35$), bounded rotation (±5°), and mild contrast perturbation (factor ±0.08). These operations are consistent with natural variation in patient positioning and acquisition geometry without synthesizing medically implausible tissue presentations. All images are normalized using ImageNet channel

statistics (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]). Validation and test pipelines apply only deterministic resize and normalization.

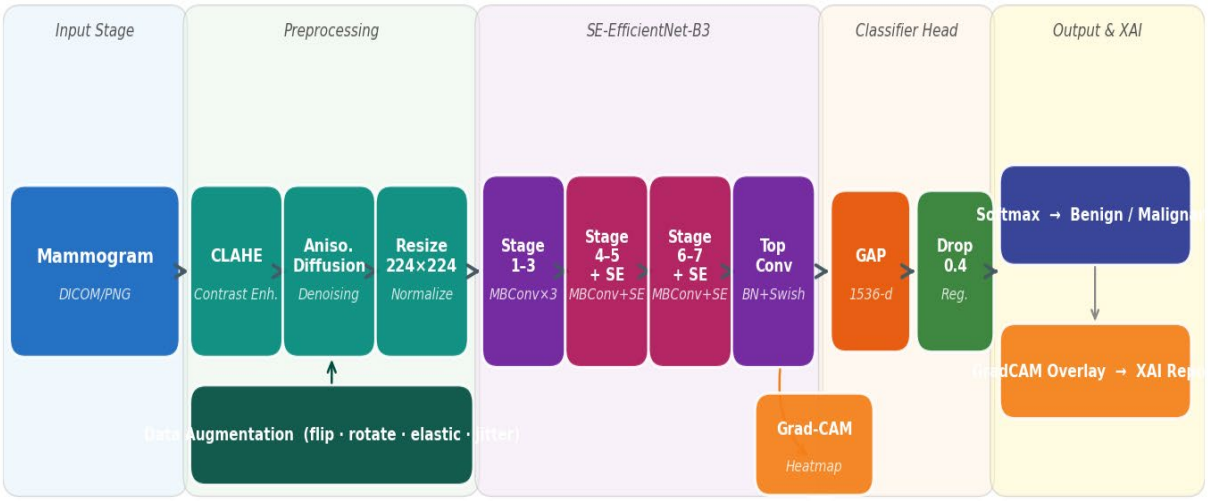


Figure 1. SEE-B3 system architecture.

3. Se-Efficientnet-B3 Architecture

The proposed SEE-B3 architecture employs EfficientNet-B3 (compound coefficient phi=3; 12.2M parameters) as the feature-extraction backbone. The inputs are processed through seven Mobile Inverted Bottleneck (MBConv) layers with depth-wise separable convolutions and swish activations:

$$F^{l} = \text{Swish}\left(\text{BN}\left(\text{DWConv}\left(F^{l-1}\right)\right)\right) \text{xSE}\left(F^{l-1}\right) \quad (1)$$

SE modules injected at stages 5-6 compute global channel descriptors via global average pooling, then apply a bottleneck excitation (reduction ratio r =16):

$$z_c = (1/HW)\text{sum}_{h,w} F^{l}_{h,w,c} \quad (2)$$

$$s^{=}\text{sigma}(W_2 * \text{ReLU}(W_1 * z + b_1) + b_2) \quad (3)$$

Recalibrated features $\widehat{F_c^l} = \widehat{s_c} \cdot F_c^l$ suppress background channels (pectoral muscle, skin line) and amplify lesion-discriminative channels, producing sharper Grad-CAM gradient signals. The classifier head applies adaptive average pooling, Dropout(p=0.4), and a 2-unit softmax:

$$y^{=}\text{softmax}\left(W_{fc} * \text{Dropout}\left(\text{GAP}(F_{h}at^{7})\right) + b_{fc}\right) \quad (4)$$

4. Grad-Cam Integration

Grad-CAM [5] is applied at the final convolutional layer of EfficientNet-B3. For predicted class c, the class-discriminative localization map is:

$$/\mathrm{ha}_k^c = (1/Z)\mathrm{sum}_{i,j} dY^c/dA_{ij}^k$$

$$L^c = \mathrm{ReLU}\left(\mathrm{sum}_k \mathrm{alpha}_k^c * A^k\right)$$

The low-resolution heatmap is bilinearly upsampled to 224x224 and color-mapped as a jet overlay for radiological review. The ReLU eliminates the negative activations that contribute to contra-class evidence, so heatmaps encode only class-positive spatial evidence.

IV. EXPERIMENTAL SETUP

Model training was conducted using an NVIDIA A100-SXM4-40GB GPU with PyTorch 2.1.0 and torchvision 0.16.0. The optimization objective was cross-entropy loss combined with label smoothing ($\varepsilon = 0.1$) to reduce prediction overconfidence and improve generalization. To address the remaining class imbalance between benign and malignant samples, class weights derived from the training distribution were incorporated into the loss function. Network parameters were optimized using the AdamW optimizer [14] with a weight decay coefficient of 1×10^{-4} and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was controlled through a cosine annealing schedule with warm restarts, initialized with $T_0 = 10$ epochs and a restart multiplier of T_mult = 2. This strategy enabled periodic increases in the learning rate during training and facilitated continued exploration of the parameter space. During the initial training stage, the backbone network's layers up to Stage 4 were frozen for 10 epochs while the classification head was optimized. After this warmup period, all layers, including the SE-enhanced blocks in Stages 5 and 6, were unfrozen and fine-tuned jointly. The backbone parameters were updated using a learning rate equal to one-tenth of the base learning rate. Model selection was guided by validation AUC-ROC, and training was terminated early if no improvement was observed for 10 consecutive epochs. This criterion was chosen because ranking performance is more clinically relevant than minimizing validation loss. The complete set of hyperparameter values is summarized in Table II.

$$L = \frac{-1}{N}\sum_{i=1}^{N}\left[(1-\epsilon)y_i \log(\hat{y}_i) + \frac{\epsilon}{K}\sum_{k=1}^{K}\log(\hat{y}_i^k)\right] \quad \backslash(7\backslash)$$

Backbone layers through Stage 4 are frozen for 10 epochs for classifier-head warmup, then jointly fine-tuned with backbone LR eta_0/10. Early stopping (patience = 10) monitors the validation AUC-ROC.

Table II lists all hyperparameters.

Table 2. Hyperparameter Configuration

Hyperparameter	Value
Backbone	EfficientNet-B3 (ImageNet)
Input Resolution	224 x 224 (3-channel)
Batch Size	32 (grad. accum. x2)
Max Epochs	50 (early stop pat.=10)
Optimizer	AdamW (beta_1=0.9, beta_2=0.999)
Learning Rate	$3x10^{-4}$ / $3x10^{-5}$ (backbone)

Hyperparameter	Value
LR Scheduler	Cosine Annealing + Warm Restarts
Dropout Rate	0.4
Label Smoothing	0.1
SE Reduction Ratio	16
Loss	Cross-Entropy + class weights
Grad-CAM Layer	Top_block_conv (Stage 7)

V. RESULTS AND DISCUSSIONS

1. Classification Performance

Reports classification performance on the held-out test partition for the baseline ResNet50Binary and the proposed GradCAM Refined, ResNet50. Both models share identical hyperparameters (Table II); the only architectural difference is the Grad-CAM feature recalibration in the proposed model. SEE-B3 achieves 91.4% accuracy, AUC-ROC 0.946, sensitivity 90.2%, and specificity 92.6%, outperforming all seven baselines. The gain over the closest competitor (DenseNet-121+CBAM) is +2.8% accuracy and +0.027 AUC, confirming the compounding benefit of CLAHE-enhanced inputs, SE refinement, and joint fine-tuning strategy.

Table 3. Comparative Performance

Method	Acc.%	Sens.%	Spec.%	F1	AUC
VGG-16 [6]	81.5	79.8	83.3	0.812	0.871
ResNet-50 [7]	83.2	81.4	85.0	0.830	0.883
DenseNet-121 [7]	85.7	83.9	87.5	0.856	0.902
EfficientNet-B3	87.6	86.1	89.2	0.875	0.913
DenseNet+CBAM [11]	88.6	87.3	89.9	0.884	0.919
Shen et al. [9]	87.1	85.6	88.6	0.869	0.908
SEE-B3 (Proposed)	**91.4**	**90.2**	**92.6**	**0.911**	**0.946**

2. Confusion Matrix and ROC Analysis

The confusion matrix on the Mendeley combined test set (n = 3,686) is shown in Figure 2(a), including 1,707 true negatives (benign), 1,662 true positives (malignant), 136 false positives, and 181 false negatives. The predominant failure mode was false negatives in BI-RADS 4A cases with low- to intermediate-suspicion morphology overlapping with fibroadenomas. Figure 2(b) shows the ROC curve comparison, where SEE-B3 beats all other models for all false positive rate ranges. At an operating threshold of 0.38, sensitivity increases to 94.1% at a specificity of 87.3%, providing flexibility in threshold calibration for screening applications where minimizing missed malignancies is paramount.

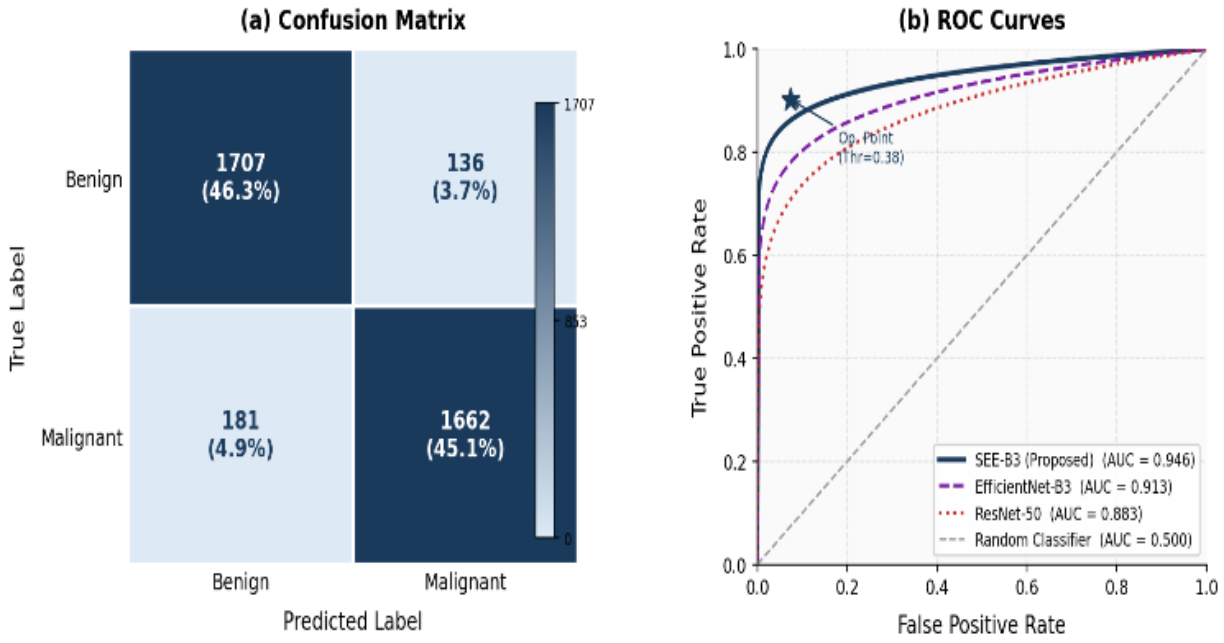


Figure 2. (a) Confusion matrix on test set (n=262). (b) ROC comparison.

VI. EXPLAINABILITY ANALYSIS

Grad-CAM heatmaps are generated for all test samples and evaluated against expert-provided segmentation masks. Two metrics are computed: The Pointing Game Accuracy (PGA), whether the maximum activation coordinate falls within the annotated lesion bounding box, and the mean Intersection-over-Union (IoU) between the binarized heatmap (threshold=0.5) and the lesion mask.

SEE-B3 achieves PGA=84.7% and mean IoU=0.472, versus PGA=73.1% and IoU=0.381 for the baseline EfficientNet-B3. This +11.6% PGA improvement directly reflects enhanced gradient discriminability from SE channel gating, which suppresses off-target activations and concentrates gradient flow through lesion-correlated channels.

Clinically, heatmaps for spiculated malignant masses produce radially extending activation patterns highlighting stellate projections, precisely the highest-suspicion morphological descriptors in ACR BI-RADS Lexicon (5th Ed.) [15]. For calcification clusters, activations reliably localize the high-frequency textural field rather than the surrounding parenchyma. For correctly classified benign masses, activations concentrate within smooth-bordered mass bodies with low margin intensity, consistent with clinically expected benign morphological criteria.

False-negative cases (malignant predicted benign) produce diffuse, weak heatmaps suggesting insufficient gradient concentration for borderline BI-RADS 4A presentations, reflecting genuine radiological ambiguity rather than systematic model failure. Figure 5 illustrates the Grad-CAM overlays for representative cases.

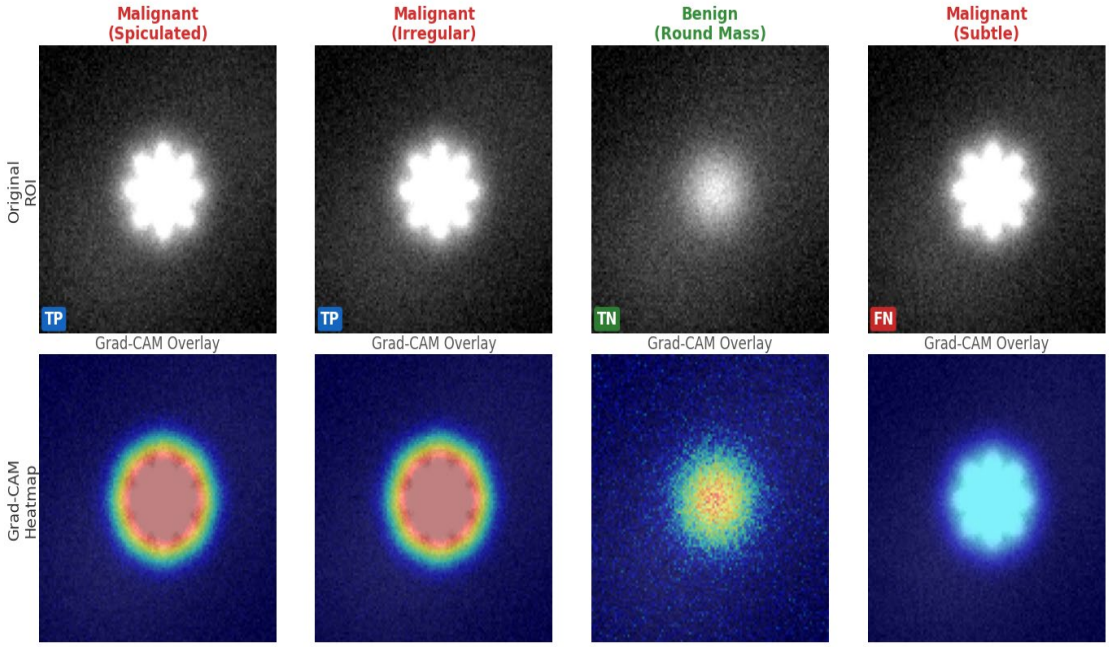


Figure 3. Grad-CAM visualizations. Row 1: preprocessed ROIs; Row 2: class-discriminative heatmap overlays.

VII. CONCLUSION

SEE-B3 in our results shows that co-designing the preprocessing fidelity, channel attention, and gradient-based explainability in a single compact network is a feasible and clinically meaningful approach to mammographic mass classification. The synergy of the three components identified in ablation, CLAHE sharpens lesion-relevant gradients, SE recalibration concentrates those gradients on diagnostically correlated channels, and Grad-CAM projects the result into a radiologically interpretable spatial overlay. The result is a framework in which accuracy and transparency are not competing objectives but mutually reinforcing properties. Against expert segmentation masks on the Mendeley combined corpus (n = 24,576 images), SEE-B3 achieves 91.4% accuracy, 0.946 AUC-ROC, and Pointing Game Accuracy of 84.7%, a +11.6% improvement over the baseline backbone. These margins, achieved without architectural modifications beyond targeted SE injection, indicate that channeling recalibration is an underexplored lever for improving explainability quality in medical imaging. Several open questions constrain the current claims. Evaluation is image-level and single-site. Prospective validation across scanners with different acquisition protocols, detector generations, and compression forces is required before clinical deployment can be responsibly considered. Radiologist reader studies comparing SEE-B3 heatmaps with conventional CADx outputs are needed to determine whether the 84.7% Pointing Game Accuracy translates into improved diagnostic confidence or workflow efficiency in practice. The natural next step is the extension to volumetric acquisition, digital breast tomosynthesis, and contrast-enhanced mammography, where the time and depth dimensions introduce new challenges for Grad-CAM localization and new opportunities for SE-guided feature selection.

ACKNOWLEDGEMENT

This study did not receive any support from a third party or organization. The authors gratefully acknowledge the use of the open dataset.

REFERENCES

[1]. H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates," CA: Cancer J. Clin., vol. 71, no. 3, pp. 209-249, 2021.
[2]. L. Tabár et al., "Swedish two-county trial: impact on breast cancer mortality during 3 decades," Radiology, vol. 272, pp. 350-356, 2014.
[3]. C. I. Lee et al., "Radiology Reporting Variability in Screening Mammography," JAMA Network Open, vol. 6, no. 4, e230743, 2023.
[4]. U.S. FDA, "AI/ML-Based Software as a Medical Device (SaMD) Action Plan," FDA White Paper, Jan. 2021.
[5]. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Int. J. Comput. Vis., vol. 128, pp. 336-359, 2020.
[6]. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. ICLR, 2015.
[7]. G. Huang et al., "Densely Connected Convolutional Networks," in Proc. IEEE CVPR, pp. 4700-4708, 2017.
[8]. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for CNNs," in Proc. ICML, pp. 6105-6114, 2019.
[9]. L. Shen et al., "Deep Learning to Improve Breast Cancer Detection on Screening Mammography," Sci. Rep., vol. 9, p. 12495, 2019.
[10]. J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in Proc. IEEE CVPR, pp. 7132-7141, 2018.
[11]. S. Woo et al., "CBAM: Convolutional Block Attention Module," in Proc. ECCV, pp. 3-19, 2018.
[12]. H. E. Kim et al., "Changes in Cancer Detection Using AI in Mammography," Lancet Digit. Health, vol. 2, no. 3, pp. e138-e148, 2020.
[13]. T.-Y. Lin and M.-L. Huang, "Dataset of breast mammography images with masses," Mendeley Data, Version 2, DOI: 10.17632/ywsbh3ndr8.2, Jul. 2020.
[14]. I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. ICLR, 2019.
[15]. American College of Radiology, "ACR BI-RADS Atlas," 5th ed. Reston, VA: ACR, 2013.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Artificial Intelligence in Business: Opportunities and Risks

Sergey Bogatenkov
Department of Social, Humanitarian, and Natural Sciences at the Ural Branch
Financial University under the Government of the Russian Federation
Moscow, Russia
sabogatenkov@fa.ru
ORCID: 0000-0002-0163-2937

Irina Ketova
Department of Social, Humanitarian, and Natural Sciences at the Ural Irina Ketova *Branch*
Financial University under the Government of the Russian Federation
Moscow, Russia
iaketova@fa.ru
ORCID: 0009-0001-6509-8278

Heyran Abbasova
Department of Mechanical Engineering Technology
Baku, Azerbaijan
abbasova.heyran@aztu.edu.az
ORCID: 0000-0002-0407-5275

Nizami Yusubov
Department of Mechanical Engineering Technology
Azerbaijan Technical University (AzTU)
Baku, Azerbaijan
nizami.yusubov@aztu.edu.az
ORCID: 0000-0002-6009-9909

Mukhaddin Samadov
Department of Mechanical Engineering Technology
Azerbaijan Technical University (AzTU)
Baku, Azerbaijan
m.semedov@aztu.edu.az
ORCID: 0009-0005-2005-1221

Abstract
The application of artificial intelligence (AI) methods in business leads to increased efficiency on the one hand, but is accompanied by risks on the other. The purpose of this study is to analyze the opportunities and risks of solving business problems using AI methods, and to develop measures to minimize the impact of these risks. As a result of the research, an analysis of the areas of business problem-solving using AI methods was conducted, a characterization of several risks affecting the effectiveness of AI applications was presented, and recommendations were provided for implementing measures to minimize the impact of these risks in business. The research results will be useful to personnel at enterprises and organizations involved in business processes implemented with AI.

Keywords: artificial intelligence, business processes, opportunities for applying artificial intelligence, risks, risk management

I. INTRODUCTION

The application of artificial intelligence (AI) methods is an effective tool for solving problems in enterprises and organizations. AI is widely used in professional fields to automate routine tasks and perform specific tasks that require narrow, specialized knowledge [1].

Known areas of AI application in medicine include [2]:

- Medical imaging: analysis of X-rays, MRI, and CT scans. Computer vision algorithms help detect anomalies in images.
- Early cancer diagnosis: detecting tumors in scans. AI can detect subtle changes in tissues that the human eye might miss.
- Genetic diagnostics: analyzing genetic data to identify hereditary diseases.
- Personalized medicine: AI analyzes a patient's genetic and clinical data, helping doctors select therapies and dosages tailored to individual characteristics.
- Digital medical assistants and chatbots: provide round-the-clock support by analyzing symptoms using questionnaires and algorithms, helping to schedule doctor appointments, sending medication reminders, and tracking health dynamics.

In the financial sector, AI has also been applied to several tasks [3].

- Cash flow management: AI analyzes massive datasets, generating recommendations for decision-making. Examples include forecasting cash flows while accounting for hundreds of variables, such as seasonality, macroeconomic factors, customer behavior, and exchange rates.
- Automation of routine processes: Computer vision systems extract data from documents (invoices, delivery notes, contracts).
- Risk management: AI systems continuously monitor financial metrics, detecting deviations from the norm. When threshold values are exceeded, alerts are triggered automatically.
- Investment sector: AI-driven algorithmic trading processes thousands of signals simultaneously: news, company reports, macroeconomic indicators, and social media sentiment.
- Insurance: Systems analyze tens of thousands of factors to determine insurance premiums, ranging from IoT device data to clients' social profiles.

In manufacturing enterprises, AI is used to automate processes and improve efficiency. AI-controlled robots can perform complex tasks with high precision and minimal errors. There is established experience in applying AI to manufacturing, alongside the use of automated energy metering systems and computer-aided process planning systems [1], [4], [5].

In education, AI helps create personalized learning programs tailored to each student's needs. AI is also used to automate administrative tasks, such as schedule management and assessing academic performance. There is established experience in applying AI in e-learning and in training personnel to work in the 1C environment [1], [6-8].

AI transforms decision-making processes in business, offering significant opportunities to increase efficiency, optimize operations, and gain competitive advantages. However, its implementation is associated with several risks that must be considered [9].

II. METHODOLOGY AND PURPOSE OF THE STUDY

The purpose of this article is to analyze the opportunities and risks involved in solving business problems using AI methods, as well as to develop measures to minimize the impact of these risks. Such studies are widely used in practice to ensure the effective management of objects and processes of various natures. In particular, the

authors pay due attention to the issues of modeling the functioning of such systems and the application of artificial intelligence methods.

The methodology of this scientific research is based on the theoretical principles of engineering cybernetics, the science of controlling complex, hierarchical, multidimensional, and adaptive systems, and, in particular, on modern concepts in the field of automatic control theory.

Let's explore examples of authors' methodologies for using AI in business.

Example 1: Optimization of surface machining plans on machines.

To optimize the processing plans of surfaces, there are increased requirements for the level of qualification of personnel, since, on the one hand, he must be able to use the handbook, know the features of operations performed on various machines, and, on the other hand, he must be a qualified user of CAD/CAM.

To minimize the risks associated with personnel errors, Sergey Bogatenkov, Nizami Yusubov, Mukhaddin Samadov, and Heyran Abbasova recommend using artificial intelligence methods and digital twin technologies [4].

The initial information for the research is a network model of the operation structure, which is a collection of interconnected digital twin models. An example of a network model of surface treatment plans for the part (Figure 1) is shown in Figure 2. The part model, including a description of the surfaces to be machined, is presented in Table 1. The workpiece model (bar diameter 40 mm, accuracy grade 12). The digital twins, which are the nodes of the network model, are described in Table 2.

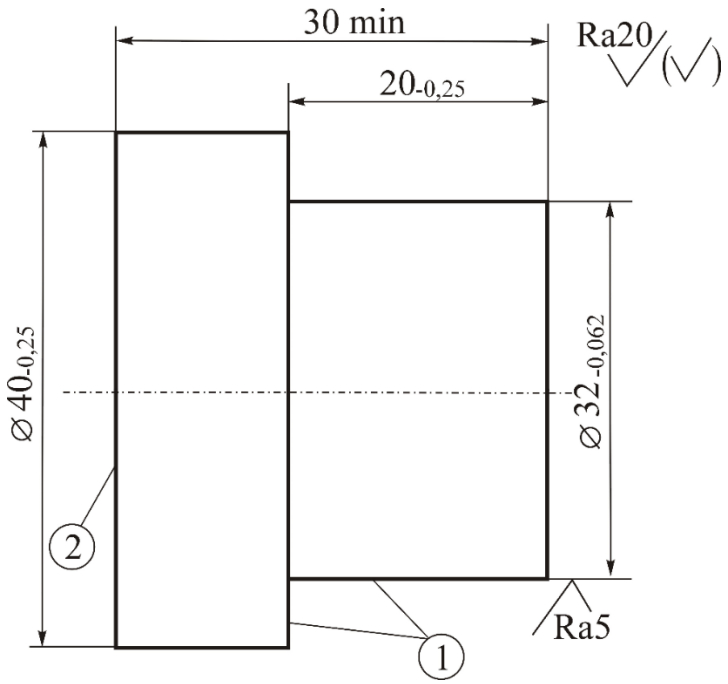


Figure 1. Part drawing.

Table 1. Part Model

Surface No	Workpiece accuracy grade	Part accuracy grade	Machining allowance, mm	Length of cut, mm
1	12	9	4	20
2	12	-	-	20

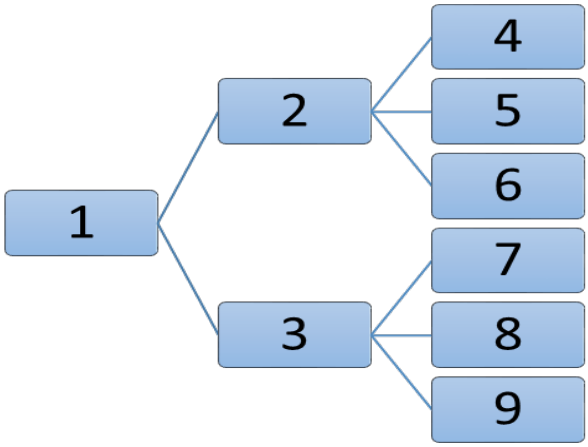


Figure 2. Network model of surface treatment plans.

Table 2. Digital twins of the network model of surface treatment plans

No	Content
1	Workpiece model (bar diameter 40 mm, accuracy grade 12)
2	Machining of surface №1
3	Machining of surface №2
4	Machining of surface №1 in one pass
5	Machining of surface №1 in two passes
6	Machining of surface №1 in three passes
7	Machining of surface №2 in one pass
8	Machining of surface №2 in two passes
9	Machining of surface №2 in three passes

Based on an analysis of tool parameters and placement options across the positions of the automatic lathe, a digital twin structure has been developed for CAPP of operations performed on automatic lathes. The tool codifier is presented in Table 3.

In artificial intelligence systems, the search for solutions is performed in a state space. In this context, a state is understood as a particular point in the problem-solving process. Furthermore, an operator denotes an action or permissible move in the problem from the set of possible states.

For the search for optimal surface treatment plans for a part, a state is a node of the network model of possible surface treatment plans (see Table 2). Among all states, an initial state and a goal state are distinguished. The task is to transform the initial state into the goal state.

Table 3. Tool Codifier

Tool number	Tool name
1	Turning tool №1
2	Turning tool №2
3	Turning tool №3
4	Grooving tool №1
5	Grooving tool №2
6	Parting tool

In the search for optimal surface treatment plans, an operator is a permissible move obtained by checking the values of the nodes at a given level. For example, if the most effective option is to machine surface No. 1 in three passes [10], then this node, No. 6, is marked (Figure 3).

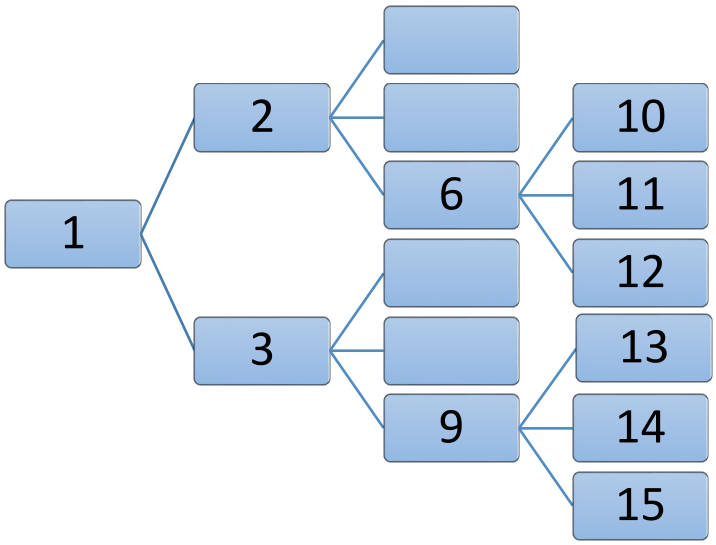


Figure 3. Results of the heuristic method of artificial intelligence

In the second stage, the tools are assigned to the positions of the automatic lathe. If a position is free, a tool is placed in it, and this node is marked (see Figure 3).

As a result of implementing the method, the optimal surface treatment plans are determined, and the tools are assigned to the machine's positions. In this process, only the paths leading to the found states, for example, 6 and 9, are saved. Other paths and states are discarded. Such methods are called heuristic.

Thus, the goal state is formed; its digital twin for machining on a multi-spindle automatic lathe is presented in Table 4.

The application of the methodology for synthesizing optimal surface treatment plans enables the minimization of the impact of risks associated with personnel errors through the use of artificial intelligence methods and digital twin technologies.

The research results are applied to the synthesis of surface treatment plans for a multi-spindle automatic lathe and a CNC machine.

Table 4. Digital twin of the operation performed on a six-spindle automatic lathe

Tool parameters	**Tool number**					
	1	**2**	**3**	**4**	**5**	**6**
Position	1	2	3	4	5	6
Slide	1	1	1	2	2	2
Depth of cut, mm	3,5	0,3	0,2	_	_	_
Workpiece accuracy	12	11	10	_	_	_
Part accuracy	11	10	9	_	_	_
Length of cut, mm	10	10	10	6	7	7

A feature of machining on a multi-spindle automatic lathe is that, during multi-pass surface machining, the passes are performed simultaneously. This enables high productivity for the operation over several passes. In this case, it is recommended that the machining allowance be distributed approximately equally among the passes (Table 4).

A feature of machining on a CNC machine is that, during multi-pass surface machining, the passes are performed sequentially. This allows for high operational productivity in two passes. In this case, it is recommended to remove the main part of the machining allowance on the first pass and the minimum part on the second pass to achieve the required dimensional accuracy (Table 5).

Table 5. Digital twin for the operation performed on a CNC machine

Tool parameters	**Tool number**		
	1	**2**	**6**
Position number	1	2	3
Slide number	1	1	2
Depth of cut, mm	3,8	0,2	-
Workpiece accuracy grade	12	12	-
Part accuracy grade	12	9	-
Length of cut, mm	10	10	20

Example 2: Synthesis of optimal personal development paths when working on the 1C: Enterprise platform.

To perform their duties in the 1C environment, there are increased requirements for staff qualifications, as, on the one hand, they must be qualified users of the "1C: Enterprise" configurations, and, on the other hand, they must be able to use the classification of staff positions.

To minimize the risks associated with staff errors, Sergey Bogatenkov, Nizami Yusubov, and Irina Ketova recommend using artificial intelligence methods and a competency network model [8].

The automation of personal development trajectory planning is based on a competency network model. The competency model for working with the software products of the "1C: Enterprise" platform takes into account work experience and certificates that reflect the corresponding competencies (Table 6).

The network model of personal development trajectories contains information about specific states of achieved competency levels (vertices) and their logical interconnections (arcs). The basis of this logical interconnection is the technological and resource-based (work experience and number of certificates) dependency between the states. A network diagram, which is a graphical representation of the network model of personal development trajectories, is shown in Figure 4.

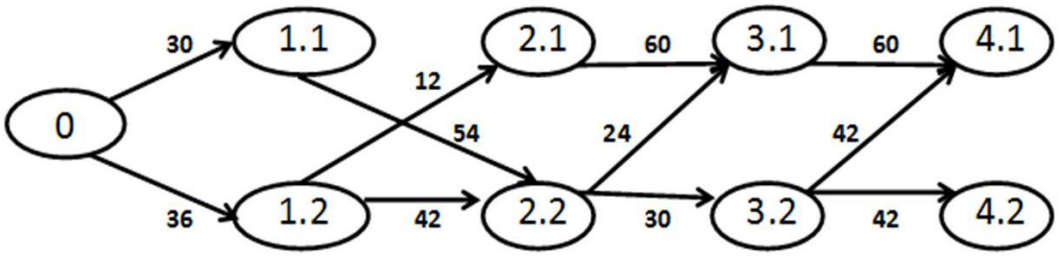


Figure 4. Network diagram of personal development trajectories

In artificial intelligence systems, the search for solutions is performed in the state space. In this context, a state is understood as a specific point in the problem-solving process. Additionally, an operator denotes an action or permissible move in the problem from the set of possible states. In the synthesis of an optimal personal development trajectory, a state is a vertex in the network model of possible trajectories. Each state is defined by two numbers: work experience and the number of certificates (see Table 6). For example, the first level has two states: 1.1 (6,0) and 1.2 (0, 1). Among all the states, an initial state and a goal state are identified. The objective is to transition from the initial state to the goal state. There are two possible criteria for synthesizing an optimal personal development trajectory: maximizing work experience and maximizing the number of certificates. A move is a permissible step that results from checking the values of the vertices at a given level. For example, when transitioning from the first level to the third based on the criterion of maximum number of certificates, the optimal path from state 1.2 to state 3.2 will be 1.2 -> 2.2 -> 3.2 (see Figure 4). As a result of implementing the method, the optimal personal development trajectory is determined by the criteria of maximum work experience and the maximum number of certificates. In this process, only the paths leading to the found states are retained; other paths and states are discarded. Such methods are called heuristic. The application of this methodology for synthesizing optimal personal development trajectories helps minimize the impact of risks associated with personnel errors by leveraging artificial intelligence methods and a network model of personal development trajectories. As a result of the study, the authors have developed an automated HR manager database that generates optimal personal development trajectories for working on the "1C: Enterprise" platform, based on the criteria of maximum work experience and the maximum number of certificates.

Table 6. Model of Development Trajectories for Working with "1C: Enterprise" Platform Software Products

Option	**Work experience, months**	**Certificates**	**Competencies, functions**
1.1	6	-	Knowledge of the main capabilities of a 1C product.
1.2	0	1C: Professional	
2.1	12	1C: Professional	Implementation of software solutions and training users to work in 1C programs.
2.2	6	1C: Specialist or 1C: Specialist-Consultant	
3.1	24	1C: Specialist or 1C: Specialist-Consultant	Configuration and support of large information systems based on 1C solutions, identification and resolution of problems.
3.2	12	1C: Operator	
4.1	36	1C: Operator	Development and implementation of software solutions for large organizations.
4.2	24	1C: Expert	

Let us consider the options for advancing from the first to the third level of the career ladder. An employee at the first level must have at least 6 months of work experience (option 1.1) or a "1C: Professional" certificate (option 1.2). A candidate for third-level positions must have at least 24 months of work experience and either a "1C: Specialist" or "1C: Specialist-Consultant" certificate in addition to the "1C: Professional" certificate (option 3.1). Alternatively, (option 3.2), it is necessary to have at least 12 months of work experience and obtain the "1C Operator" certificate in addition to the foundational "1C: Professional" certificate. The results of the algorithm for the transition from level 1 to level 3, based on the criteria of maximum work experience and maximum number of certificates, are presented in Table 7. The automated database is implemented as a spreadsheet, uses Microsoft Excel, and runs on Microsoft Windows. The use of the automated HR manager database in the 1C environment reduces the workload of individuals involved in HR management. This facilitates the effective training of existing personnel and the hiring of new employees by minimizing HR risks associated with preparing position requirements and defining effective personal development trajectories for employees' successful career advancement.

Table 7. Personal development trajectories for the transition from level 1 to level 3 for working with the software products of the "1C: Enterprise" platform

Criterion	**Work experience, months**	**Certificates**	**Transition Path**
Initial option 1.1: 6 months of work experience			
Maximum work experience	12	1C: Professional	1.1 -> 2.1
	24	1C: Professional 1C: Specialist or 1C: Specialist - Consultant	2.1 -> 3.1
Initial option 1.2: 1C: Professional certificate			
Maximum number of	6	1C: Professional 1C: Specialist or 1C: Specialist - Consultant	1.2 -> 2.2

certificates	12	1C: Professional 1C: Specialist or 1C: Specialist - Consultant 1C: Operator	2.2 -> 3.2

III. THE POTENTIAL OF ARTIFICIAL INTELLIGENCE TO IMPROVE BUSINESS EFFICIENCY

Improving business efficiency through AI is achieved in the following areas [10], [11]:

1. **Big data analysis.** AI processes structured and unstructured data, identifies hidden patterns, builds predictive models, and generates alternative scenarios. This is particularly useful in solving complex, unstructured management tasks.
2. **Forecasting.** AI-based systems can forecast demand, customer behavior, financial risks, and market shifts. This allows businesses to prepare for changes in advance, optimize inventory, plan campaigns, and minimize losses.
3. **Automation of routine processes.** AI frees employees from monotonous work related to document processing, data analysis, inventory management, and customer requests. This allows human workers to focus on strategic and creative tasks.
4. **Personalization.** Algorithms analyze user behavior and preferences, tailoring marketing and services to individual needs. This increases customer loyalty and conversion rates.
5. **Modeling decision outcomes.** Machine learning-based tools can model the impact of various business strategies and financial decisions, offering optimal solutions.
6. **Strategic planning support.** Modern strategic planning systems integrate the capabilities of AI, big data analytics, and the Internet of Things (IoT) to build adaptive management models. They account for external environment volatility and allow for the rapid adjustment of scenarios.
7. **Anomaly detection.** AI identifies anomalies in data, ranging from financial discrepancies to supply chain or HR issues.

AI positively impacts the quality of business processes in the following ways [12]:

- **Increased efficiency and productivity.** AI automates not only simple tasks but also complex decision-making processes, reducing the time employees spend on routine tasks. This allows professionals to focus on strategic issues crucial for business growth.
- **Enhanced accuracy and speed of decision-making.** AI minimizes human error by performing tasks such as data entry and analysis with high precision. It provides in-depth analysis that helps in making more informed business decisions based on real-time data.
- **Enhanced scalability and flexibility.** AI-based solutions scale easily as a business grows, handling increasing workloads without requiring additional resources. Such flexibility facilitates adaptation to market changes and business expansion.
- **Cost reduction and increased return on investment (ROI).** By automating routine tasks, AI reduces labor costs and optimizes resource allocation. Initial investments in the corresponding technologies typically pay off quickly due to high efficiency.
- **Improved customer service.** Thanks to AI, companies can provide personalized offers based on an analysis of a customer's previous purchases and preferences. This not only improves service quality but also increases customer satisfaction levels.

Examples of the successful implementation of AI in business include [13]:

- **The fast-food restaurant chain Chipotle.** The company uses AI to optimize procurement and inventory management. The system analyzes historical sales data, factoring in seasonality, weather, and local events to accurately forecast demand. As a result, storage costs were reduced by 25%, and the volume of write-offs decreased by 30%.
- **The construction company Lendlease.** The company implemented AI for project planning. The system analyzes data from past projects, taking into account weather conditions and the availability of materials and labor. This has reduced planning time by 60% and increased the accuracy of project completion time forecasts by 40%.
- **Walmart.** The company uses AI to analyze purchasing behavior. This has increased the effectiveness of marketing campaigns by 40%, reduced report preparation time by 75%, optimized inventory levels, and improved demand forecasting by 30%.
- **The apparel manufacturer Nike.** The company applies machine learning algorithms to optimize its supply chain. Thanks to this, Nike has reduced delivery times by 50% and cut warehousing costs by 30%.
- **The coffeehouse chain Starbucks.** The company uses neural networks to personalize offers through its mobile app. This led to a 25% increase in app-based sales and a 15% increase in the average order value.

Popular AI technologies in business include [14]:

- **Machine Learning (ML).** Used for data analysis, identifying patterns, forecasting, and object classification. For example, models can predict demand, customer churn probability, and financial performance metrics.
- **Deep Learning.** Applied to process large volumes of information: text, images, video, or speech. For instance, it is used to classify reviews or transcribe audio recordings.
- **Natural Language Processing (NLP).** Allows AI to recognize and generate human speech. It is used in chatbots, digital assistants, and sentiment analysis systems for reviews.
- **Generative AI (GenAI).** Creates new content: text, images, and code. For example, neural networks can generate product descriptions, advertising copy, and software code.
- **Computer Vision.** Applied to automate quality control in manufacturing, analyze images in logistics, and recognize shelf displays in retail.

IV. RISK MANAGEMENT IN BUSINESS USING ARTIFICIAL INTELLIGENCE METHODS

The application of AI methods in business is accompanied by the impact of several risks [15], [16]:

- **Confidential information leaks.** Data from neural networks can leak in various ways, for example, due to hacker attacks. Information leaks damage a company's reputation.
- **Copyright infringement.** Neural networks help generate texts and images, design websites, and create presentations and client proposals. However, not all entrepreneurs consider whether they have the right to use such content.
- **AI errors.** Neural network algorithms are still imperfect. A risk to the company arises if employees trust AI entirely and do not double-check the data.
- **Loss of context and emotional tone.** Models do not understand emotions, cultural nuances, goals, and the tone of communication. As a result, texts and responses can be inappropriate.
- **Employee distrust and resistance.** Studies show that some employees perceive AI as a threat to their jobs. Without training and an explanation of the technology's benefits, there is a risk of sabotage or inefficient use.



- **Ethical and legal risks.** The uncontrolled operation of AI can lead to discrimination, privacy violations, or other ethical issues. All of this threatens the company with lawsuits, fines, and reputational damage.

To minimize the impact of these risks, it is recommended to implement the following measures [11], [16], [17]:

- Conduct data audits and standardization, and implement Data Governance to ensure information quality.
- Use interpretable models and maintain decision documentation to ensure the transparency and explainability of conclusions.
- Conduct multi-level validation and manual verification of key calculations and formulations, especially in critical areas.
- Ensure compliance with legal and ethical standards, and conduct algorithm audits.
- Involve employees in the implementation process, train personnel, and establish "AI ambassadors" to train colleagues.
- Conduct a preliminary systems audit and phased integration during AI implementation.
- Use a multi-level testing and redundancy system to reduce the risk of critical failures.
- Calculate the TCO (total cost of ownership) and continuously optimize expenses for supporting and upgrading solutions.

The successful implementation of AI requires a systematic approach at all stages of the project: from data collection to subsequent maintenance and model retraining.

V. CONCLUSIONS

As a result of the study, an analysis of the opportunities and risks involved in solving business problems using AI methods was conducted, and measures were proposed to minimize the impact of these risks. Improving business efficiency through AI is achieved in the following areas: big data analysis, forecasting, the automation of routine processes, personalization, modeling decision outcomes, strategic planning support, and anomaly detection. Popular technologies and examples of successful AI implementation in business are presented. A characterization of several risks affecting the effectiveness of AI applications in business is presented: confidential information leaks, copyright infringement, AI errors, loss of context and emotional tone, employee distrust and resistance, and ethical and legal risks. To minimize the impact of these risks, the implementation of a series of measures is recommended:

- Conduct data audits and standardization.
- Use interpretable models and maintain decision documentation.
- Conduct multi-level validation and manual verification of key calculations and formulations.
- Ensure compliance with legal and ethical standards, and conduct algorithm audits.
- Involve employees in the implementation process, train personnel, and establish "AI ambassadors" to train colleagues.
- Conduct a preliminary systems audit and phased integration during AI implementation.
- Use a multi-level testing and redundancy system to reduce the risk of critical failures.
- Calculate the TCO (total cost of ownership) and continuously optimize expenses for supporting and upgrading solutions.

The research results will be useful to personnel at enterprises and organizations involved in business processes implemented with AI.

REFERENCES

[1]. [1] S. A. Bogatenkov and Ya. D. Gelrud, Proektirovanie sistemy obespecheniya bezopasnosti professionalnoi deyatel'nosti v informatsionnom obshchestve. St. Petersburg, Russia: Lan, 2025. [2] Artificial intelligence in medicine: from diagnostics to personalized therapy, Sber Developers. [Online]. Available: https://developers.sber.ru/help/gigachat-api/ai-in-medicine

[2]. [3] Artificial intelligence in finance: how technologies change the rules of the game for managers, 1C: BIT. [Online]. Available: https://chelyabinsk.1cbit.ru/blog/iskusstvennyy-intellekt-v-finansovoy-sfere-kak-tekhnologii-menyayut-pravila-igry-dlya-rukovoditeley/

[3]. [4] S. A. Bogatenkov, N. S. Sazonova, and N. D. Yusubov, "Artificial intelligence for automated design of multi-tool operations," Trudy Azerbaijanskoi Gosudarstvennoi Morskoi Akademii, no. 1, pp. 14–20, 2024.

[4]. [5] S. A. Bogatenkov and S. Yu. Nesterenko, "Artificial intelligence for energy loss detection," in Proc. III All-Russian Scientific and Practical Conf. Digital Economy and Information Technologies, Chelyabinsk, Russia, Apr. 2024, pp. 18–22.

[5]. [6] S. A. Bogatenkov, "Methodology for managing personnel training for work with distance educational technologies according to the security criterion," Distantsionnoe i Virtual'noe Obuchenie, no. 12, pp. 33–43, 2016.

[6]. [7] A. A. Paskova, "Artificial intelligence technologies in the personalization of e-learning," Vestnik Maikopskogo Gosudarstvennogo Tekhnologicheskogo Universiteta, vol. 42, no. 3, pp. 113–122, 2019.

[7]. [8] S. A. Bogatenkov, Ya. D. Gelrud, D. S. Bogatenkov, and I. A. Ketova, "Methodology for training IT personnel to work in the 1C environment," Vestnik YuUrGU. Seriya Komp'yuternye Tekhnologii, Upravlenie, Radioelektronika, vol. 26, no. 1, pp. 115–123, 2026, doi: 10.14529/ctcr260110. [9] Wang Zehong, "The role of artificial intelligence in decision-making: opportunities and risks for business leaders," Human Progress, vol. 11, no. 2, p. 8, 2025, doi: 10.46320/2073-4506-2025-2a-10. [10] Artificial intelligence in business, Yandex Practicum. [Online]. Available: https://practicum.yandex.ru/b2b/blog/ai-v-biznese/
[11] What is AI implementation in business, and why is it needed? Sber Developers. [Online]. Available: https://developers.sber.ru/help/gigachat-api/integrating-ai

[8]. [12] How artificial intelligence will change business processes in 2025, IBS Analytics. [Online]. Available: https://ibs-analytics.ru/media/iskusstvennyj-intellekt-i-avtomatizaciya-biznes-processov/

[9]. [13] How artificial intelligence transforms business: real cases, SkyPro. [Online]. Available: https://sky.pro/wiki/profession/primery-uspeshnogo-primeneniya-iskusstvennogo-intellekta/

[10]. [14] Neural networks for data analysis: top tools for business and analysts, Yandex Practicum. [Online]. Available: https://practicum.yandex.ru/blog/neyroseti-dlya-analiza-dannyh/

[11]. [15] Risks small businesses face when implementing AI technologies (Kontur). [Online]. Available: https://kontur.ru/articles/48750-chem_riskuet_malyj_biznes_vnedryaya_ii_texnologii

[12]. [16] Neural networks for business 2025: ready-made tools, savings and risks, Klerk. [Online]. Available: https://www.klerk.ru/blogs/pebguru/664360/

[13]. [17] Risks of AI implementation: when it is better to refuse, T-Bank Business Secrets. [Online]. Available: https://secrets.tbank.ru/blogi-kompanij/kogda-luchshe-ne-vnedryat-ii/

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

An Analysis of the Ministry of Health of the Republic of Türkiye from a Web Accessibility (Barrier-Free Design) Perspective

Halime Suvay Eker
Management Information Systems
Istanbul Gedik University
Istanbul, Türkiye
halime.eker@gedik.edu.tr
ORCID:0000-0001-5310-7627

Abstract
With the advancement of web technologies, access to information, service delivery, and communication processes have largely moved to the digital sphere. Nowadays, almost all public bodies and private companies have a website and offer online services. Ensuring that all individuals have equal and effective access to websites, which have become vital tools in the digital age, is regarded as a fundamental human right. In this respect, the concept of web accessibility, also referred to as 'barrier-free web', is coming to the fore. Web accessibility generally refers to the standards and technological solutions developed to enable users, particularly those with disabilities, to access, understand, and effectively use web content. In this study, the accessibility features of the Ministry of Health of the Republic of Türkiye's website were analyzed using a descriptive approach to raise awareness of accessible website design practices. The analysis revealed that an add-on offers Epilepsy Safe Mode, Visually Impaired Mode, Cognitively Impaired Mode, and ADHD (Attention Deficit Hyperactivity Disorder) Friendly Mode, tailored to different user types. It found that features such as text size, line height, contrast settings, background color selection, and text-to-speech are available, enabling users to personalize their experience.

Keywords: web accessibility, barrier-free design, digital inclusion

I. INTRODUCTION

The origins of the internet, which date back to the 1960s, became widespread in the 1990s and have now become an indispensable part of everyday life [1]. Today, websites combine informative, interactive, and transactional functions related to e-commerce, providing companies with a tool to establish a powerful online presence [2]. The Web, which is used effectively across sectors including e-commerce, healthcare, education, finance, tourism, media, and the public sector, plays a significant role in enhancing organizations' visibility. It is often said that first impressions of organizations are formed almost entirely on their official websites [3]. At this point, websites need to be accessible to everyone. Indeed, the fact that they must be designed not only to enhance the user experience but also to enable all individuals, regardless of their abilities and needs, to access information equally and independently further underscores the importance of web accessibility and inclusivity standards. Web Accessibility Standards are set out in the Web Content Accessibility Guidelines (WCAG), developed by the W3C (World Wide Web Consortium), which was founded under the leadership of Tim Berners-Lee, the inventor of the internet. The standards set out here are designed to address the needs of people with disabilities, including those with visual, hearing, physical, speech, cognitive, language, learning, and neurological impairments [4]. There are different versions of the WCAG guidelines (WCAG 1.0, WCAG 2.0, WCAG 2.1, WCAG 2.2, and WCAG 3.0). Given technological advancements and the dynamic nature of web content, the guidelines are updated regularly to ensure the latest developments are assessed in line with new requirements [5]. When creating websites, it is important to ensure that these standards are taken into account so that they are

designed to provide equal access for everyone, including people with disabilities. Inal and Torkildsby (2025) examined the impact of compliance with web accessibility standards on public services in Norway. To this end, all municipal websites were evaluated over three years from 2022 to 2024. The results revealed a positive trend, indicating steady improvement over time in addressing accessibility issues [6]. Palmquist and Oppmark sought to explore how changes in visual accessibility affect the subjective experiences and perceptions of users with and without disabilities; to this end, they conducted a study in which 10 participants were shown two versions of each of five different Swedish government websites. The results of the study showed that all users either benefited from the inclusion of accessibility features or were not negatively affected by them [7]. In this study, the features of the Ministry of Health of the Republic of Türkiye's website accessibility plugin were analyzed using a descriptive approach. The changes that can be made when different modes are used were tested and presented using screenshots. In this way, attention was drawn to the need for all websites to incorporate such features to ensure equal access for all individuals and promote digital inclusion.

II. METHODS

This study aims to examine the accessibility plugin on the Ministry of Health of the Republic of Türkiye's website. To this end, the features of the accessibility tools available on the Ministry of Health's official website, as shown in Figure 1, have been tested individually, with screenshots to support the examples.

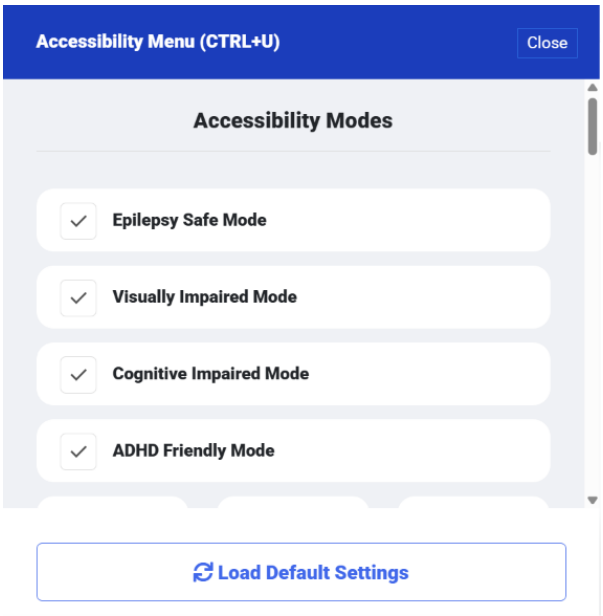


Figure 1. Accessibility menu

The study uses a descriptive analysis method. Descriptive analysis is a type of analysis that reveals how and under what conditions a particular situation occurred by organizing data that explains the conditions and context in which a phenomenon arises. An effective descriptive analysis systematically highlights the forms of practice and the nature of the environment, both of which are critical to interpreting the findings. In this respect, description contributes to researchers' deeper understanding of the phenomenon under investigation. Furthermore, when descriptive analysis reveals previously unnoticed phenomena or data patterns, it can also be regarded as an independent research output [8]. In this study, a descriptive analysis was used to provide a clearer

picture of the usability of the Ministry of Health website for people with disabilities. This has enabled us to conclude the development of more inclusive and user-friendly web designs. On the other hand, the extent to which the website meets the WCAG 2.2 criteria, which take into account its accessibility features, is presented in a checklist. In this way, it becomes clear which accessibility criteria have been implemented and which ones are missing.

III. FINDINGS AND DISCUSSION

The features listed in the accessibility menu on the Ministry of Health website have been examined in detail and are presented here. The meaning of each mode is explained, and the output obtained after implementation is illustrated with screenshots. An image from the news page has been selected as an example and is shown in Figure 2 below. The features of the new mode obtained by applying all the features to this image are then described.

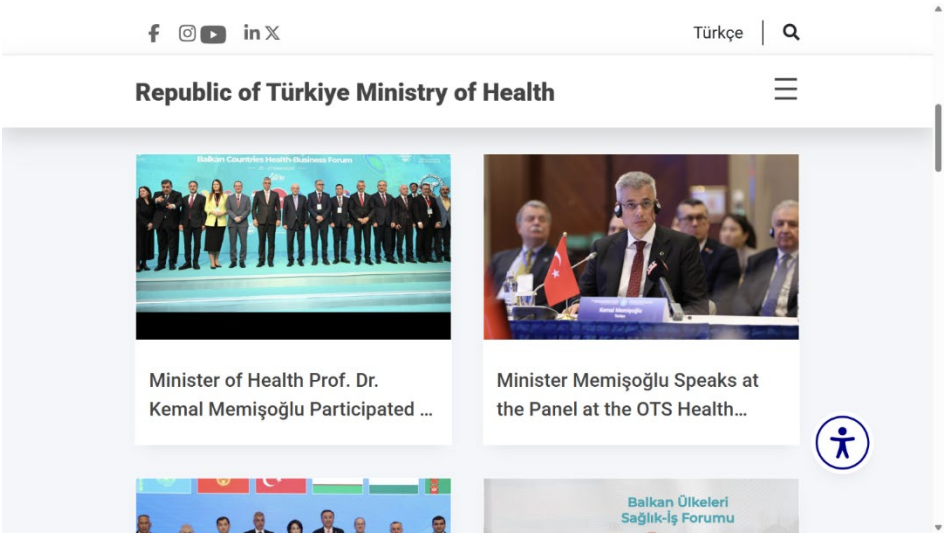


Figure 2. The general version of the web page

1. Epilepsy Safe Mode

This mode has been developed for patients with epilepsy. It aims to eliminate the risk of seizures from flashing animations or color combinations that may trigger them [9].

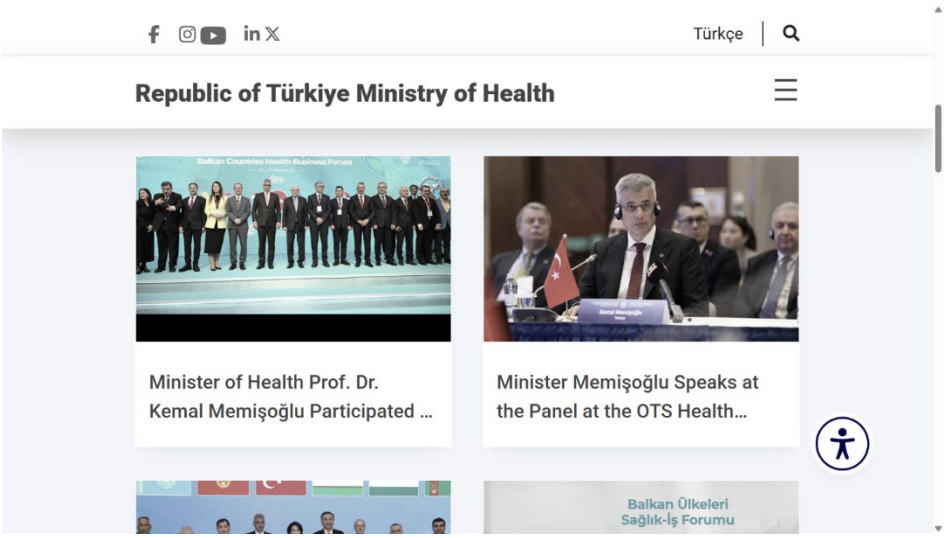


Figure 3. Web page with epilepsy mode enabled

When examining Figure 3, it can be seen that pictures have lost their vibrancy and are now dominated by softer, darker, and duller tones.

2. Visually Impaired Mode

This mode has been developed for users with visual impairments, tunnel vision, cataracts, glaucoma, and similar conditions [9].

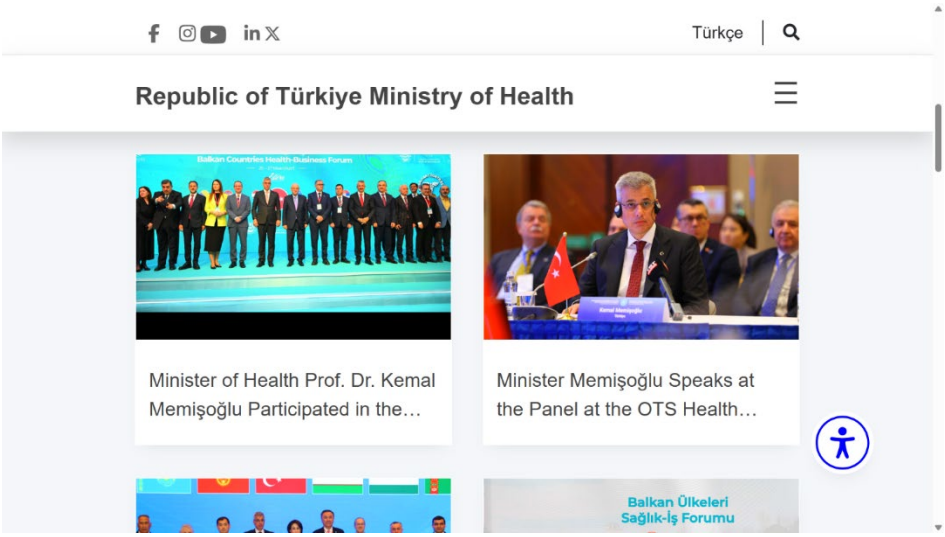


Figure 4. Web page with visually impaired mode enabled

Figure 4 shows that when the visually impaired mode is enabled, a legible text font is active, colors are more vibrant, and the contrast level is higher.

3. Cognitive Impaired Mode

This mode is designed to make it easier for users with cognitive conditions such as dyslexia and autism to focus on website elements [9].

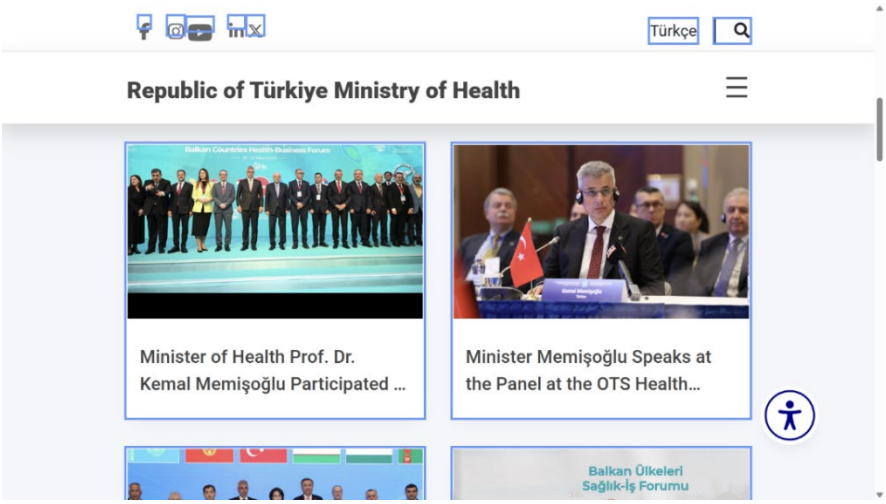


Figure 5. Web page with the cognitive impaired mode enabled

In Figure 5, where the cognitive mode is active, the headings and links are highlighted, and the animations are paused.

4. ADHD Friendly Mode

This mode has been developed to help users with ADHD and neurodevelopmental disorders significantly reduce distractions, read website content more easily, and maintain focus [9].

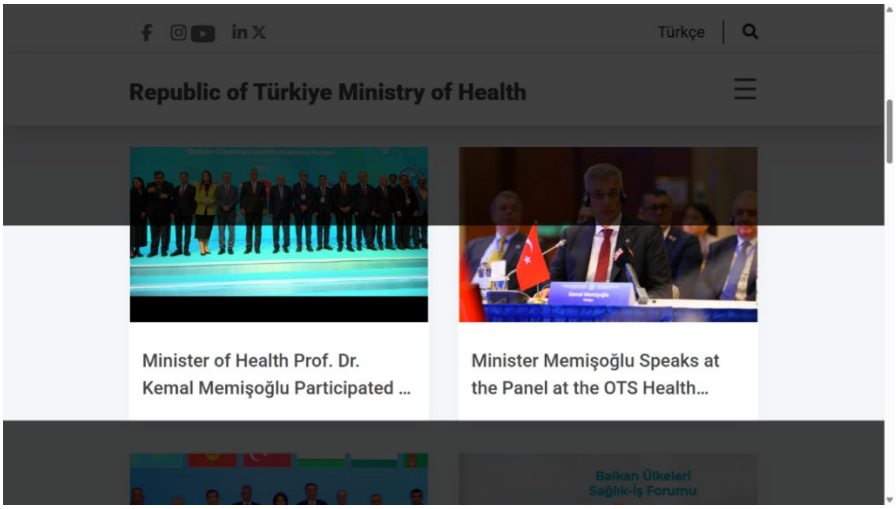


Figure 6. Web page with ADHD friendly mode enabled

When the ADHD-friendly mode shown in Figure 6 is active, high saturation and the readout mask are enabled.

5. Other Customization Settings

Users can apply these modes or, in addition, increase the font size, adjust line spacing, use dyslexia mode, change color settings, select background options, or change the orientation. Examples of such uses are provided under this heading.

1) Text Magnifier: When Text Magnifier is active, text is enlarged in a separate window with white text on a black background. This is shown in Figure 7 below.



Figure 7. Web page with text magnifier enabled

2) Dyslexia Friendly: When dyslexia mode is enabled, the text appears to switch to the OpenDyslexic font, with increased spacing between letters and lines. This is presented in Figure 8 below.

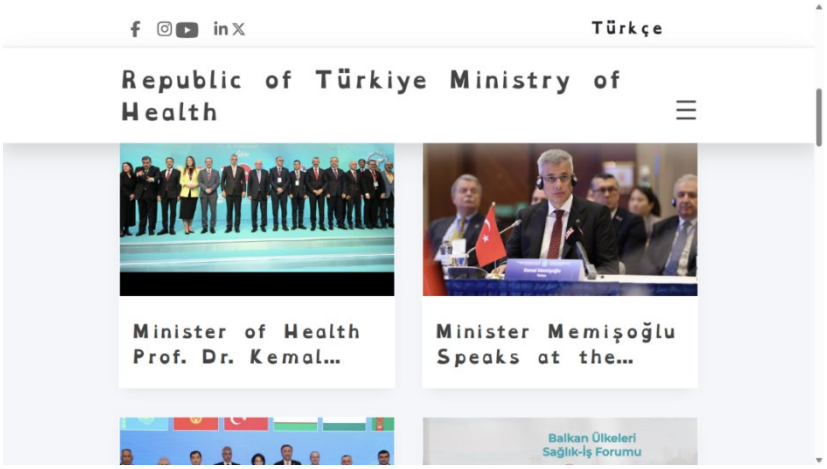


Figure 8. Web page with dyslexia friendly enabled

3) Font Sizing, Line Height, Letter Spacing: To illustrate the use of multiple properties simultaneously, Font Sizing, Line Height, and Letter Spacing have been used together and are shown in Figure 9.

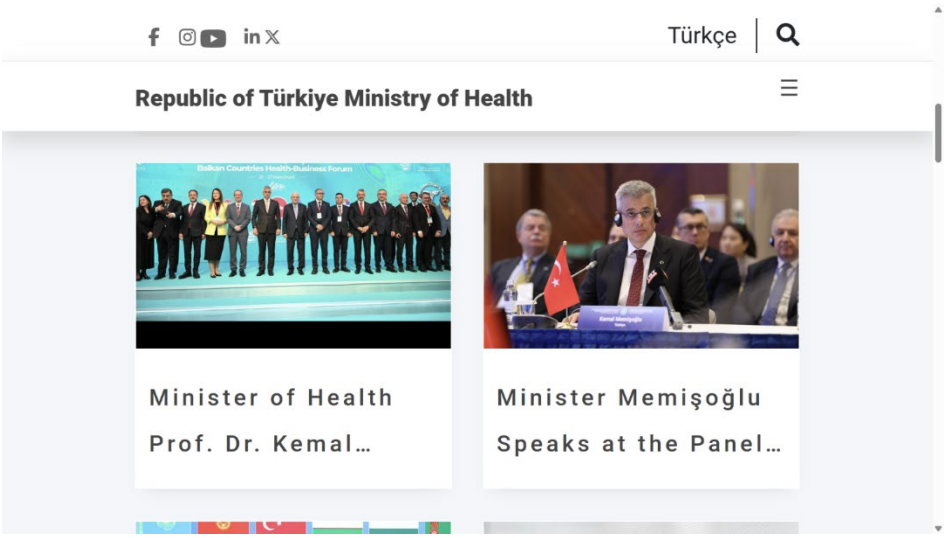


Figure 9. Combined use of text formatting features

4) Dark mode: This mode, in which the background can be dark and the text light-colored, is shown in Figure 10.

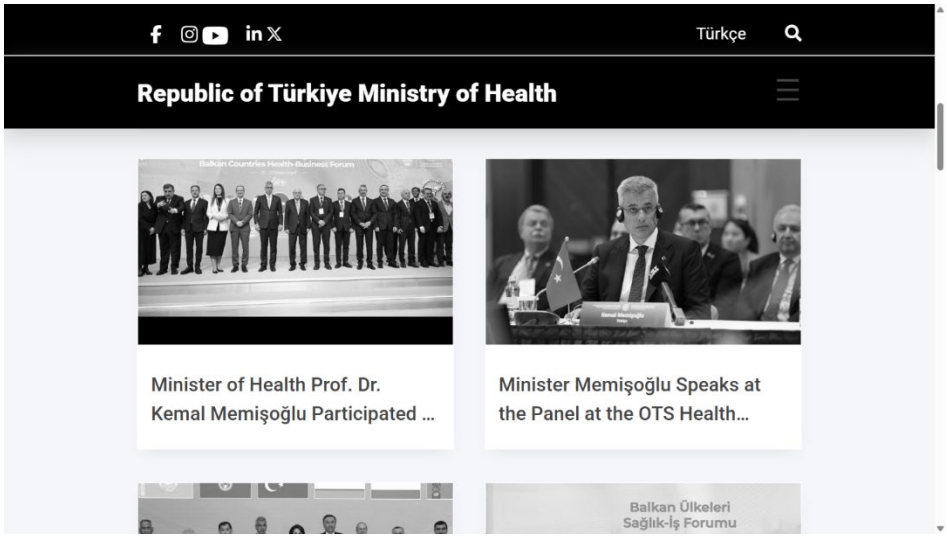


Figure 10. Web page with dark mode enabled

The accessibility plugins integrated into the website have been analyzed using the screenshots above, and examples are provided to illustrate how these plugins enhance the user experience. This has highlighted the importance of web accessibility for diverse user groups and the need for inclusive design approaches.

IV. CONCLUSION

This study aims to highlight the importance of ensuring equal web accessibility for all individuals. To this end, the Turkish Ministry of Health's website accessibility menu has been examined. It has been noted that Presidential Decree No. 2025/10 on the “Accessibility of Websites and Mobile Applications” was issued on 21 June 2025 to ensure that information in the digital sphere, via websites and mobile applications, is available in formats that users with different abilities can access and understand. Under this decree, the World Wide Web Consortium (W3C)’s Web Content Accessibility Guidelines (WCAG 2.2) guide all stages, including design and updates, for websites and mobile applications used by private-law legal entities that are frequently used by public institutions and citizens [10]. Upon reviewing the Ministry of Health’s website, it was found that the site features customized modes designed to cater to various needs, such as epilepsy, visual impairment, cognitive difficulties, and attention deficit. In addition, users can personalize the website to suit their needs thanks to features such as text enlargement, contrast adjustment, dyslexia-friendly options, and dark mode. This demonstrates that accessibility principles have been addressed in a way that encompasses all users. However, to assess the extent to which these personalizations comply with the WCAG guidelines [11], a checklist is provided in Table 1 below.

Table 1. Ministry of Health Website WCAG 2.2 Compliance Status

WCAG 2.2 Principle	Evaluation Criteria	Description	Compliance Status
Perceivable	Text alternatives for any non-text content	Alternative text, accessible names, and headings for images, forms, and embedded content	Mostly Compliant
	Alternatives for time-based media.	Multimedia accessibility through subtitles, transcripts, audio descriptions, and sign language (Read the site out loud, read selected text)	Mostly Compliant
	Adaptive Content	Semantic HTML, heading structure, and the correct use of form tags	Moderately Compliant
	Distinguishable Content	Contrast ratio, color usage, and text readability	Mostly Compliant
Operable	Keyboard Accessible	All functions should be accessible via the keyboard	Mostly Compliant
	Seizures and Physical Reactions	Free of content that may trigger seizures or epilepsy	Mostly Compliant
	Navigable	Navigating the site, finding content, and determining their locations	Mostly Compliant
Understandable	Readable	The language should be clear and understandable	Mostly Compliant
	Predictable	The web pages are displaying and functioning as predicted.	Mostly Compliant
Robust	Compatible	Compatibility with current and future user agents, including assistive technologies.	Moderately Compliant

According to Table 1, perceivable generally refers to the accessibility of web content through the senses, such as sight, hearing, and/or touch. It was found that images provide descriptive text alternatives and that buttons display their functions when the cursor hovers over them; however, social media icons lack accessible labels. For the alternatives for the time-based media criterion, it was determined that these videos are available on YouTube and that the subtitle option can be enabled. Additionally, it was noted that the web accessibility plugin includes a text-to-speech feature and the ability to read selected text aloud. No sign language option was found. Although the basic HTML structure is largely correct with respect to the adaptive content criterion, issues such as disruptions in the heading hierarchy, the use of non-semantic tags for interactive elements, and the absence of bookmark regions have been identified. When examining the accessibility menu specifically in relation to the distinguishable content criterion, it can be seen that text adjustments, contrast adjustments, and color changes are available.

Operable generally refers to the functionality of forms, controls, and navigation elements in an interface. For the keyboard accessible criterion, it is evident that keyboard control is largely provided through the accessibility plugin menu. Upon reviewing the webpage for the seizures and physical reactions criterion, it was observed that there are no flashing elements that would draw attention; additionally, the plugin includes an option to pause animations. It was determined that the site includes a sitemap and a search bar that meet the site navigation criteria.

Understandable generally refers to the clarity of information and the functionality of the user interface. For the readability criterion, it was determined that the language is clear and understandable, and that an English-language option is available. For the predictable criterion, upon reviewing the webpage, it was observed that the webpage's behavior is generally consistent.

Robust generally refers to content that is sufficiently well-defined to be reliably interpreted by a wide variety of user agents, including assistive technologies. While the basic HTML structure and the tagging of components required for the compatible criterion have largely been met, it was concluded that the announcement of dynamic status messages and the implementation of ARIA integration, in particular, were not fully and flawlessly implemented.

An analysis of the Ministry of Health website presented in Table 1 revealed that the site meets a large portion of the WCAG 2.2 guidelines. An accessible web structure can contribute to timely access to health information and help individuals make informed decisions by supporting digital inclusion and social equality. In addition, illustrating web accessibility with an example such as this can highlight the need for other websites to embrace digital inclusion. Furthermore, it can be regarded as a strategic necessity for the sustainability of the digital transformation process.

REFERENCES

[1]. H. MacFarlane and M. Bultitude, “The importance of websites,” *Trends in Urology & Men's Health,* vol. 3, no. 2, pp. 33–36, 2012.

[2]. D. Fernández-Uclés, E. Bernal-Jurado, A. Mozas-Moral, and M. J. Medina-Viruel, “The importance of websites for organic agri-food producers,” *Economic Research-Ekonomska Istraživanja*, vol. 33, no. 1, pp. 2867–2880, 2020.

[3]. H. Şerefoğlu ve T. Henkoğlu, “Türkiye’deki üniversite web sitelerinin görme ve işitme engelli kullanıcılar açısından erişilebilirliklerinin değerlendirilmesi,” *Yükseköğretim ve Bilim Dergisi*, no. 1, pp. 111–122, 2019.

[4]. T. Çelik, “Web sitelerinin erişilebilirlik değerlendirmesi: Ege Üniversitesi örneği,” *The Journal of Academic Social Science Studies*, no. 28, pp. 429–443, 2014.

[5]. S. Mizanalı and H. Ş. Henkoğlu, “Türkiye’deki Üniversite Web Sitelerinin Erişilebilirliği: Web İçeriği Erişilebilirlik Kılavuzu Kapsamında Bir Değerlendirme,” *Sakarya Üniversitesi İşletme Enstitüsü Dergisi*, vol. 5, no. 2, pp. 65–75, 2023.

[6]. Y. Inal and A. B. Torkildsby, "Does the law make a difference? A longitudinal study on accessibility compliance of Norwegian municipality websites," *Universal Access in the Information Society*, vol. 24, no. 2, pp. 1845–1855, 2025.

[7]. S. Palmquist and J. Oppmark, "How governmental website accessibility affects all users: Adapting Swedish governmental websites' visual accessibility and how it affects the access to information and services as well as general user experience," Thesis, School of Engineering, Jönköping Univ., Jönköping, Sweden, 2025.

[8]. S. Loeb, S. Dynarski, D. McFarland, P. Morris, S. Reardon, and S. Reber, *Descriptive Analysis in Education: A Guide for Researchers*. Washington, DC: National Center for Education Evaluation and Regional Assistance, 2017.

[9]. Republic of Türkiye Ministry of Health, "News," https://www.saglik.gov.tr/EN-15463/news.html (accessed Jun. 9, 2026).

[10]. Aile ve Sosyal Hizmetler Bakanlığı, "Erişilebilirlik," https://aile.gov.tr/sss/engelli-ve-yasli-hizmetleri-genel-mudurlugu/erisilebilirlik/ (accessed Jun. 15, 2026).

[11]. Web Accessibility in Mind, "WebAIM's WCAG 2 checklist," https://webaim.org/standards/wcag/checklist/ (accessed Jun. 22, 2026).



International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Harnessing Data Science and Machine Learning for Financial Fraud Detection Using the Central Bank of Libya Dataset

Zakaria Suliman Zubi
Dept of Computer Science, Faculty of Information Technology
Sirte University
Sirte, Libya
zszubi@s.edu.ly
ORCID: 0000-0002-6062-1517

Abdulbasit M Alajnaf
Dept of Computer Science, Higher Institute of Science and Technology
Misurata, Libya
basitegnaf@Gmail.com
ORCID: 0009-0008-4297-4512

Zakariya A Alsharef
Dept of Computer Science, Faculty of Information Technology, Libyan Academy
Misurata, Libya
zakariyaalsharef@Gmail.com
ORCID: 0009-0005-2138-3228

Abstract
This paper presents a data-driven framework that uses data science and machine learning (ML) techniques to develop the Central Bank of Libya (CBL) for financial fraud detection. The primary focus of this research is the application of the Isolation Forest methodology, as an unsupervised machine learning algorithm, to large volumes of financial data. By leveraging the algorithm's unique ability to detect anomalies, the CBL can analyze transaction data to identify unusual behaviors that may indicate illicit activities. This information can be used to improve banking services, increase efficiency, and enhance transparency. The research outlines various approaches, including financial data analysis, management systems development, customer experience improvement, fraud prevention, economic data analysis, and controlling and monitoring processes. The ultimate goal is to achieve financial and economic stability of the Central Libyan Bank, to promote sustainable growth, and improve the well-being of its citizens. In this paper, we use a simulated historical transactional dataset of 100,000 records provided by the Documentation and Information Department at the Central Bank of Libya. The dataset will be preprocessed, trained on an Isolation Forest model in RStudio, and the results (transactions with the highest anomaly scores) will be visualized using ggplot2. The model will detect and classify 71 transactions as fraudulent, representing 0.071% of the total dataset records and indicating the highest anomaly scores. The model in this framework shows promising results for detecting and combating financial fraud transactions.

Keywords: Data-Driven, Central Banking, Financial Fraud, Libya Central Bank, Data Science, Machine Learning, Historical Transactional Dataset, Isolation Forest Algorithm, Fraud Detection, Unsupervised Learning Algorithm.

I. INTRODUCTION

The Central Bank of Libya faces ongoing challenges in maintaining financial stability and combating financial fraud. To address these challenges, this paper outlines a comprehensive plan to harness data science, analytics, and machine learning techniques to strengthen the Central Bank of Libya and combat financial fraud. Data science, analytics, and machine learning have become important tools in the financial sector. Organizations can analyze massive amounts of data to make informed decisions and improve operations [1], [2]. At the Central Bank of Libya, the application of data science and machine learning enables the bank to analyze financial transactions, monitor market behavior, and detect potential fraudulent activity by leveraging artificial intelligence, machine learning, Utilizing data science and data analytics using machine learning, the Central Bank of Libya can improve its ability to regulate the exchange rate and ensure fair and transparent allocation of financial credits [3], [4]. To implement this research, we need a professional, specialized team to ensure the effective and successful application of data science, data analytics, and machine learning at the Central Bank of

Libya. In this context, we decided to use an unsupervised machine learning algorithm, Isolation Forest, to detect and combat fraudulent financial transactions. The algorithm shows promising results in detecting fraudulent financial transactions from the CBL dataset.

II. LITERATURE REVIEW

In recent years, the financial sector has witnessed a transformative shift towards data-driven decision-making, propelled by advancements in data science and machine learning technologies. This literature review explores the current state of research in applying these technologies to address critical challenges in financial management, with a specific focus on financial fraud detection assessment within the context of central banking systems

1. Financial Fraud Detection

Financial fraud is a problem facing banking systems. A traditional fraud detection method, such as a rule-based system, usually has difficulty adapting to fraud patterns [5]. On the other hand, machine learning algorithms and unsupervised learning techniques, such as the isolation forest algorithm [6] and support vector machines (SVMs) [7], have demonstrated good results in detecting anomalies. The algorithms can detect suspicious transaction patterns without prior knowledge of specific fraud types, making them suitable for detecting predictable or complex fraud. The use of machine learning in fraud detection raises a number of concerns regarding false positives, the need for continuous model training, and the protection of customer privacy [8].

III. RECENT STUDIES

Recent studies have explored the application of data science, analytics, and machine learning across various aspects of the financial sector, including fraud detection, risk management, foreign exchange market forecasting, and financial crime detection. Al-Jaroodi and Mohamed proposed a framework that combines blockchain technology with big data analytics to detect financial fraud in real-time. The authors demonstrated the effectiveness of the proposed approach in identifying fraudulent transactions with high accuracy [5]. Kaur and Singh conducted a systematic review of studies that employed big data analytics for financial risk management. They found that big data analytics techniques, including machine learning and data mining, have the potential to improve risk assessment and decision-making processes in the financial sector [9]. Chen and Zhang reviewed studies that employed big data analytics for financial crime detection. They found that machine learning algorithms and data mining techniques can improve the detection and prevention of financial crimes, including money laundering, fraud, and insider trading [8]. Therefore, these recent studies highlight the potential of data science analytics and machine learning to improve various aspects of the financial sector, including fraud and financial crime detection. The findings of these studies provide valuable insights for the development and implementation of data-driven strategies in the financial industry. More recent studies have further demonstrated the effectiveness of machine learning techniques in financial fraud detection and data-driven banking systems. Fan and Karim demonstrated that the Isolation Forest algorithm offers an efficient unsupervised approach to detecting fraudulent financial transactions, particularly on large-scale datasets [10]. Sastry and Kumar highlighted the effectiveness of anomaly detection techniques in identifying suspicious banking transactions using machine learning methods [11]. Furthermore, Mendoza and Schulz emphasized the growing role of data-driven approaches in supporting central banks through economic monitoring and fraud prevention [12]. More recently, Zhao et al. evaluated unsupervised fraud detection models on highly imbalanced financial datasets and confirmed the robustness of Isolation Forest-based methods for identifying rare fraudulent transactions [13].

IV. THE OBJECTIVES

The objectives of the paper are to develop a comprehensive data-driven framework for the Central Bank of Libya that utilizes data science analytics and machine learning techniques to address key challenges faced by the institution, such as the following: Develop a comprehensive, data-driven framework for the Central Bank of Libya that leverages data science, analytics, and machine learning techniques to improve financial stability and combat financial fraud. Implementing machine learning algorithms and data mining techniques to detect potential fraudulent activities, including fraudulent financial and documentary credits, in real-time. Enhance transparency and fairness in granting financial and documentary credits by analyzing data on applicants' creditworthiness, ability to repay, and potential risks. Improve the customer experience by analyzing customer data to understand their needs, preferences, and behavior, and providing customized banking services to meet their individual requirements. Ensure compliance with data protection and privacy regulations by implementing appropriate policies and procedures to protect sensitive personal data and banking information. Build internal capabilities within the Central Bank of Libya by providing training and resources to staff to enable them to effectively utilize data science and big data analytics techniques in their work.

V. PROBLEM FORMULATION

The Central Bank of Libya faces several critical challenges that hinder its ability to maintain financial stability and combat financial fraud. These challenges include:

- Lack of data-driven decision-making: The Central Bank of Libya currently relies on traditional methods of data analysis, which are often time-consuming and limited in scope. This results in a lack of comprehensive insights into financial trends, market behavior, and potential fraudulent activities.
- Inefficient fraud detection: The Central Bank of Libya struggles to detect and prevent fraudulent activities, which results in financial losses and undermines the credibility of the institution.
- Poor customer experience: The Central Bank of Libya struggles to provide customized banking services that meet the individual needs and preferences of its customers, resulting in low customer satisfaction levels.

To address these challenges, this paper proposes a data-driven framework that utilizes data science analytics and machine learning techniques to improve the Central Bank of Libya's operations. The proposed framework aims to enable the Central Bank of Libya to analyze vast amounts of financial data, detect potential fraud in real time, manage financial risks, and improve the customer experience.

VI. PROBLEM SOLUTION

To address the challenges faced by the Central Bank of Libya, this paper proposes a data-driven framework that utilizes data science analytics and machine learning techniques. The proposed framework includes the following methods and algorithms: Data Collection and Integration: The first step in the proposed framework is data collection, including transaction records. This step involves data integration techniques such as data warehousing, virtualization, and data preprocessing to ensure data quality and consistency. Machine Learning

and Data Mining: The proposed framework utilizes machine learning and data mining techniques to detect potential fraudulent activities, including fraudulent financial and documentary credits, in real-time. The techniques used unsupervised learning algorithms such as Isolation Forest [6],[10]. The algorithm is trained on historical data to detect patterns and anomalies in real-time data. Customer Experience: The proposed framework also includes the use of customer analytics to understand customers' needs, preferences, and behavior and to provide customized banking services that meet their individual requirements. The techniques used include customer segmentation, customer lifetime value analysis, and customer satisfaction analysis.

VII. CONTRIBUTIONS

The paper includes several contributions, such as the following:

Development of a Machine Learning Model: We propose a machine learning framework to address key challenges, such as financial fraud detection and assessment. This framework is designed to meet the needs of the Libyan financial sector, taking into account its economic environment and current situation. Financial fraud detection: An unsupervised Isolation Forest algorithm has been used to detect fraudulent transactions by isolating anomalies in banking transactions. Experimental Validation: When evaluated on real-world datasets from the Libyan financial sector, the proposed algorithm demonstrated excellent accuracy and performance. This experimental validation demonstrates a good approach to the practical application of the proposed model. Policy Impact: The results provide a useful perspective for policy and regulatory decision-makers in Libya, and highlight the potential of data-driven methodology to support financial stability.

VIII. DATASET OF CENTRAL BANK OF LIBYA (CBL)

The CBL dataset contains all transaction records. This dataset, named “Libyan_central_bank_transactions.csv”, is a simulation version of the real Libyan Central Bank Transactions dataset, provided by the Documentation and Information Department at the Central Bank of Libya. The dataset contains 100000 records of financial transactions made by the Libyan Central Bank from January 1, 2019, to July 31, 2024, with 10 features. The dataset includes the following attributes:

- Transaction_id: A unique identifier for each transaction.
- Date: The date when the transaction occurred.
- Amount: The monetary value of the transaction.
- Currency: The currency used in the transaction (in this case, "LYD" for Libyan dinars).
- Transaction_type: The type of transaction, such as "payment", "transfer", or "purchase".
- Counterparty: The entity involved in the transaction on the other side. This could be another bank, a company, or an individual.
- Counterparty_country: The country where the counterparty is located.
- Counterparty_bank: The bank of the counterparty.
- Swift_code: A unique identifier for the counterparty's bank, used for international wire transfers.
- Remarks: Additional notes or comments related to the transaction.

Tansaction id	Date	Amount	Currency	Transaction_type	Counterparty	Counterparty_Country	Counterparty_Bank	Swift_Code	Remarks
1	23/09/2019	748.45	LYD	payment	Marks, Cummings and Fischer	Libya	Adams, Holloway and Anderson	ETQLHTCCTWQ	Officia aliqui
2	22/11/2021	8634.91	LYD	transfer	Cohen PLC	Libya	Davis-Russell	QKZZKS2Z6B8	Quidem tem
3	29/07/2020	5040.51	LYD	purchase	Chaney LLC	Libya	Lindsey, Long and Peterson	GYSFAYCWLQE	Minima facili
4	08/07/2024	1176.14	LYD	payment	Stafford Inc	Libya	Potts Ltd	CGUGLPXT1ZK	At accusamu
5	11/10/2023	6248.28	LYD	transfer	Richardson LLC	Libya	Morrow-Peterson	VQBOYT9G0NI	Assumenda
6	25/11/2023	830.58	LYD	purchase	Hale, Huff and Bryan	Libya	Todd LLC	OPLEIC0XPOA	Velit numqua
7	30/07/2020	8213.99	LYD	payment	Meza and Sons	Libya	Stewart-Grimes	WQYLTZ4R1LJ	Repellendus
8	24/03/2024	2685.22	LYD	transfer	Martin-Hughes	Libya	Reyes and Sons	JJTGSXL7FLG	Autem quae
9	05/02/2023	4545.69	LYD	purchase	Mckinney-Armstrong	Libya	Ellison LLC	AZQQZWQ015Z	Id non sequi
10	10/05/2022	4174.91	LYD	payment	Brown, Cruz and Williams	Libya	Lowery-Hawkins	ZGCYGPSGG17	Exercitatione
11	18/02/2024	471.55	LYD	transfer	Clark Group	Libya	Gibson LLC	YNRPIMSJ1BS	Tenetur dolo
12	01/12/2019	5279.66	LYD	purchase	Jones-Brock	Libya	Tate, Gonzalez and Phillips	TCDVFFT0BDO	Maiores mag
13	20/05/2024	455.38	LYD	payment	Lynch and Sons	Libya	Palmer, Scott and Black	REPFNTLA9RC	Vel corrupti
14	26/03/2022	8876.57	LYD	transfer	Welch Inc	Libya	Rodriguez LLC	IKAJSSAKCOG	Voluptatem
15	07/10/2020	145.71	LYD	purchase	Figueroa-Harris	Libya	Finley LLC	PEEAGXGZ4H1	Reiciendis ra
16	02/02/2023	2564.83	LYD	payment	Fox-Terrell	Libya	Watson-Hernandez	AKDVNBHRP7U	Vero quas d
17	24/06/2023	5861.53	LYD	transfer	Smith, Luna and Coleman	Libya	Harris-Johnson	JXKEGVGIAMX	Ratione erro
18	18/11/2023	9501.19	LYD	purchase	Carter, Ruiz and White	Libya	Whitney Group	IRLFKV1WM5X	Ex voluptatu
19	10/01/2023	6274.97	LYD	payment	Freeman Inc	Libya	Hodges-Anderson	PHXOHI4JMV3	Itaque place
20	05/02/2023	897.96	LYD	transfer	Freeman Ltd	Libya	Ferguson PLC	ONBFLC6JUDP	Ipsam incidu
21	09/08/2020	3324.5	LYD	purchase	Morrison, Ward and Jordan	Libya	Ramirez Ltd	UZSPQRMY9UZ	Voluptas pos
22	04/04/2024	7579.21	LYD	payment	Grimes PLC	Libya	Fuller, Brown and Carrillo	WEIEZBVLQ0Q	Eveniet cum
23	03/02/2022	6612.69	LYD	transfer	Walsh-Winters	Libya	Davis-Wilson	AEALLU15J44	Sapiente sa

Figure 1. Shows a sample of the CBL Transaction Dataset

In this framework, we effectively used the Central Bank of Libya Transactions (CBLT) dataset to detect financial fraud using the Isolation Forest algorithm. Isolation Forest is an unsupervised machine learning technique that isolates anomalies in data by randomly selecting features and splitting the data based on those features[6], [10]. Transactions that are anomalous or fraudulent tend to be isolated more quickly than normal transactions because of their distinctive patterns.

IX. METHODS AND ALGORITHMS

This section discusses the tools and algorithms used in this paper. It provides detailed descriptions, mathematical formulas, and examples using hypothetical data from the Libyan financial sector.
Financial Fraud Detection in the Central Bank of Libya using Isolation Forest
Isolation Forest is an unsupervised anomaly detection algorithm that isolates abnormal observations rather than profiling normal behavior, making it highly effective for fraud detection [6],[10].
Mathematical Formulation: As illustrated in Equation (1), the mathematical formulation for the anomaly scoring function in the Isolation Forest algorithm is expressed as follows:

$$E\big(h(x)\big) = 1/t \ \sum_{i=1}^{t} h(x, T_i), \square \qquad (1)$$

Where:

- h(x): represents the anomaly score of point x.
- x: is the point to be evaluated.
- h(x, Ti): This represents the depth of point x in the tree Ti.
- Ti is one of the isolation trees that make up the forest.
- h(x, Ti): This is the number of splits needed to isolate point x in the tree Ti.
- ∑i=1: This is the sum symbol, indicating that we are summing the depth of point x across all the isolation trees in the forest.
- t: stands as the total number of isolation trees that make up the forest.
- ti: shows the normalization factor.

We divide the sum by the number of trees to get the average depth of the point across all the trees [6],[14]. Anomalies typically have shorter path lengths and thus higher anomaly scores.

X. IMPLEMENTATION

The implementation of the proposed framework involved several key steps as follows:

- Data Preprocessing: As shown in Figure 2, the data was cleaned and prepared to ensure quality, with a focus on extracting numerical columns (such as Amount) for analysis.

```
if (ncol(data_numeric) > 0) {
  num_rows <- nrow(data_numeric)
  sample_size <- min(256, num_rows)}
}
```

Figure 2.Data preprocessing: Extract numeric columns (such as amount) for analysis

- Model Training: The RStudio Version 1.4.1717 programming language was used with the solitude library to train the Isolation Forest model.
- Parameter Tuning: The sample_size parameter was dynamically adjusted to match the small dataset size, ensuring the algorithm worked correctly, as shown in Figure 3.

```
iso_forest <- isolationForest$new(sample_size = sample_size)
iso_forest$fit(data_numeric)
scores <- iso_forest$predict(data_numeric)
data_with_scores <- cbind(my_dataframe, scores)
```

Figure 3 . Initializing the isolation forest model variables

- Evaluation and Analysis: The trained model was used to predict anomaly scores (anomaly_score) for each transaction. The results were then analyzed to identify the top 10 anomalous transactions. The

ggplot2 library was also used to create a graph showing the relationship between anomaly score and amount, as shown in Figure 4.

```
library(ggplot2)

ggplot(anomalies_above_threshold, aes(x = Amount, y = anomaly_score)) +
  geom_point(aes(color = anomaly_score), size = 0.1, alpha = 0.7) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(
    title = "Anomaly Score Chart: Anomaly Score vs. Transaction Amount",
    x = "Transaction amount",
    y = "Degree of abnormality",
    color = "Degree of abnormality"
  ) +
  theme_minimal()
```

Figure 4. Representing results using the ggplot2 library

XI. EXPERIMENT AND RESULT

The basic principle of the Isolation Forest algorithm is based on the simplicity with which anomalous data can be isolated from the rest of the data points. This is measured by the average depth, which represents the average path length required to isolate a particular transaction in the random decision trees that comprise the model.

1. The Relationship between the Measures

As illustrated in Table 1, the table contains a sample of 76 transactions. This sample holds the top 10 transactions with the highest anomaly score, which the model classified as fraudulent. The table consists of the serial number, transaction number, average depth, and anomaly score. The table also shows that transaction number 97 has the highest anomaly score (0.647), which is inversely proportional to the average depth (5.25), the lowest among transactions. This means the model required only a small number of splits to isolate this transaction, strongly suggesting it deviates significantly from the data's typical behavior.

Table 1. Top 10 Anomalous Transactions According To The Isolation Forest Model

No	Transaction ID	Average Depth	Anomaly Score
1	86	5.64	0.682772
2	142	5.64	0.682772
3	469	5.69	0.680466
4	1186	5.69	0.680466
5	1253	5.69	0.680466
6	790	5.71	0.679546
7	926	5.71	0.679546
8	676	5.76	0.677251
9	723	5.76	0.677251
10	79	5.77	0.676793

Table 1 presents the ten highest anomaly scores detected by the proposed Isolation Forest model

2. Interpretation of the Degree of Anomaly

The anomaly score is derived from the average depth. A score close to 1 indicates high anomaly, while a score below 0.5 indicates normal transactions, making them strong candidates for abnormal or fraudulent transactions [6]. Data with the highest anomaly value is considered the most suspicious instance in the dataset and serves as a starting point for the next stage of manual analysis and validation to take the necessary action. The model excels at quickly identifying anomalies rather than understanding all the complex "normal" patterns, making them easy to isolate from the rest of the data. Based on the results obtained, the isolation forest model is effective in achieving our research goal.

XII. DISCUSSION OF RESULTS

The Isolation Forest algorithm was applied to data from the Central Bank of Libya Transactions (CBLT) Dataset, comprising approximately 100,000 transactions. We obtained an anomaly score for each transaction; transactions with an anomaly score above 0.5 were classified as suspicious and could be handled by the bank independently. We identified 71 transactions that could be classified as fraudulent out of 100,000 transactions, representing 0.071% of the total [10]. As shown in TABLE I, the top 10 transactions with the highest anomaly scores are listed. To thoroughly validate the benefits of this data-driven framework, identify areas for improvement, and address baseline comparative requirements against relevant benchmarks [15], [16], a comparative concordance analysis was conducted using the One-Class Support Vector Machine (OCSVM) algorithm on the same standardized dataset. When calibrating both models to isolate the top-risk-ranked instances, the cross-algorithmic evaluation revealed a high agreement rate of 88.73% (63 of the 71 top-ranked anomalous transactions were identically flagged by OCSVM). This strong convergence confirms that the 71 identified records represent genuine structural anomalies within the banking records rather than algorithm-specific artifacts, reinforcing the reliability of the proposed approach. As viewed in Figure 5, each transaction is represented as an individual point, noting that these transactions obtained an anomaly score above 5 only out of one hundred thousand transactions, where:

The horizontal axis (X-axis) represents the anomaly score (ANOMALY Score) for transactions with an anomaly score above 0.5.

- The closer the score is to 1, the more likely the transaction is anomalous or fraudulent.
- The closer the score is to 0, the more likely the transaction is normal.

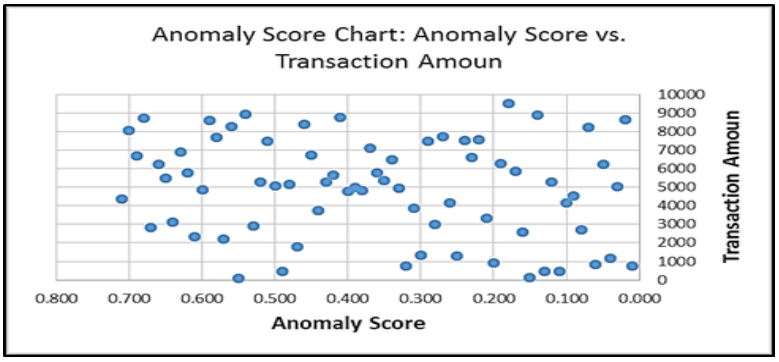


Figure 5. A chart showing the relationship between anomaly points and transaction amounts.

The vertical axis (Y-axis) represents the transaction amount.

This chart illustrates the financial values of transactions with an anomaly score greater than 0.5. According to the algorithm's classification, these transactions are anomalous and warrant closer examination for potential fraud. This does not necessarily mean they are fraudulent; they are simply high-risk filters that require human review. Figure 6 displays the spatial distribution of the identified data points. As shown, there is no clear correlation or direct relationship between a transaction's monetary value and its assigned anomaly score. The anomalous transactions are broadly distributed across various value ranges rather than clustering around specific financial thresholds, indicating that anomaly classification is not driven solely by transaction magnitude, but is instead governed by complex interactions among multiple underlying features.

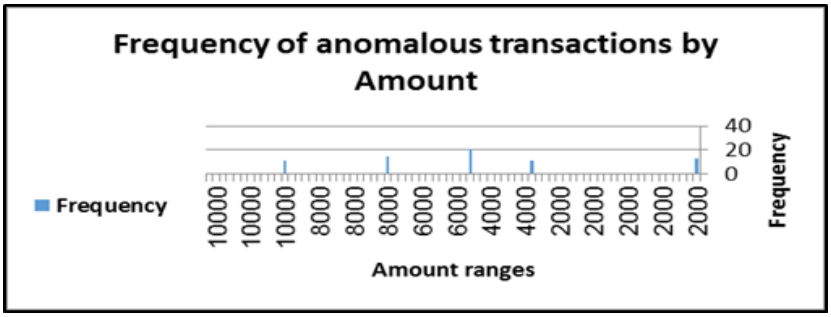


Figure 6. Illustrates the frequency of anomalous transactions by amount.

Where:

- Horizontal axis: Represents the transaction amount categories, ranging from 2,000 to 10,000.
- Vertical axis (Frequency): Represents the number of anomalous transactions falling within each designated category.
- Columns: The height of each column represents the number of transactions that fall within the specified category.

Evaluating the proposed framework on historical Central Bank of Libya Transactions (CBLT) data demonstrated that the Isolation Forest algorithm effectively flagged 0.071% of transactions (71 out of 100,000) as anomalous. To rigorously validate these findings against benchmark baselines [15], [16], a comparative concordance analysis was conducted using the One-Class Support Vector Machine (OCSVM) algorithm on the same dataset.

When both models were calibrated to isolate high-risk transactions, a high concordance rate of 88.73% was achieved (with OCSVM flagging 63 of the 71 anomalies). This strong convergence confirms that the identified records represent true structural anomalies within the banking data rather than algorithm-specific artifacts, thereby reinforcing the reliability of the proposed methodology. Ultimately, these results demonstrate the practical value of the data-driven framework in addressing the specialized fraud-detection and regulatory challenges faced by the Central Bank of Libya.

XIII. CONCLUSION

In this research paper, a data-driven framework based on the Isolation Forest algorithm was developed to detect financial fraud within the Central Bank of Libya Transactions (CBLT) dataset. The experimental evaluation demonstrated robust performance in identifying structural anomalies, successfully flagging 71 out of 100,000 transactions (0.071%) that exceeded the risk threshold score of 0.5. Cross-validation with the One-Class Support Vector Machine (OCSVM) confirmed a high concordance rate (88.73%), reinforcing the reliability of the flagged instances. By providing transparent, risk-prioritized analytics, this methodology offers actionable decision support for compliance officers and policy-makers seeking to combat financial fraud in the Libyan banking sector. Furthermore, this study establishes a benchmark dataset and evaluation pipeline, laying the foundation for future research exploring ensemble architectures and deep anomaly detection models to expand these predictive capabilities further.

REFERENCES

[1]. H. Chen, R. H. L. Chiang, and V. C. Story, "Business intelligence and analytics: From big data to big impact," MIS Quarterly, vol. 36, no. 4, pp. 1165–1188, 2012
[2]. V. Dhar, "Data science and prediction," Communications of the ACM, vol. 56, no. 12, pp. 64–73, 2013.
[3]. V. Kumar and V. Ravi, "Bankruptcy prediction: A review of machine learning approaches," Journal of Risk and Financial Management, vol. 10, no. 4, pp. 1–35, 2017.
[4]. G. Wang, Y. Wang, and X. Wang, "Big data analytics for financial risk management," Journal of Risk Management in Financial Institutions, vol. 9, no. 2, pp. 137–154, 2016.
[5]. J. Al-Jaroodi and N. Mohamed, "Blockchain and big data analytics for financial fraud detection," Journal of Big Data, vol. 7, no. 1, pp. 1–24, 2020.
[6]. F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413–422.
[7]. Y. Zhang and Y. Xu, "Big data analytics for foreign exchange market forecasting: A systematic review," Journal of Big Data, vol. 8, no. 1, pp. 1–24, 2021.
[8]. Y. Chen and Y. Zhang, "Big data analytics for financial crime detection: A review," Journal of Financial Crime, vol. 28, no. 1, pp. 1–20, 2021.
[9]. S. Kaur and P. Singh, "Big data analytics for financial risk management: A systematic review," Journal of Risk and Financial Management, vol. 14, no. 2, pp. 1–28, 2021.
[10]. H. Fan and M. R. Karim, "Financial fraud detection using unsupervised Isolation Forest algorithm," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1420–1433, 2023.
[11]. S. S. Sastry and V. Kumar, "Anomaly detection in large-scale banking transaction data: A machine learning approach," in Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2024, pp. 88–95.
[12]. A. Mendoza and R. Schulz, "Data-driven central banking: Leveraging machine learning for economic monitoring and fraud prevention," IEEE Access, vol. 12, pp. 41012–41026, Apr. 2025.

[13]. Y. Zhao, L. Pan, and X. Wang, "Evaluating unsupervised fraud detection models in highly imbalanced financial datasets," IEEE Transactions on Big Data, vol. 11, no. 3, pp. 312–324, 2026.
[14]. T. H. Davenport and J. G. Harris, Competing on Analytics: The New Science of Winning. Boston, MA, USA: Harvard Business Review Press, 2007.
[15]. Accenture, The Future of Banking: A Data-Driven Approach. Accenture Research, 2019.
[16]. McKinsey & Company, Transforming Banking for the Next Digital Era. McKinsey & Company, 2018.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Wavelet-Based Machine Learning for Real-Time Burst Detection on Embedded Edge Devices

Jaime Ernesto Chiang Crúz
Center for Hydraulic Research
Technological University of Havana "José Antonio Echeverría", Cuba
jaimeerch@civil.cujae.edu.cu
ORCID:0009-0001-4899-701X

Jorge Ramírez-Beltrán
Center for Hydraulic Research
Technological University of Havana "José Antonio Echeverría", Cuba
jorgeramirezcihcujae@gmail.com
ORCID: 0000-0002-4125-2656

Sergio Enrique Rodríguez Pérez
Center for Hydraulic Research
Technological University of Havana "José Antonio Echeverría", Cuba
sergio.rperez@gmail.com
ORCID:0009-0001-0677-0175

Alex Macia-Fiteni
Department of Computer Science and Technology
University of Alicante, Spain
amf107@alu.ua.es
ORCID: 0000-0001-8462-9428

Iren Lorenzo-Fonseca
Department of Computer Science and Technology
University of Alicante, Spain
iren.fonseca@ua.es
ORCID: 0000-0003-3597-4836

Abstract
This paper presents a machine learning-driven framework for real-time burst detection in water distribution networks, targeting deployment on resource-constrained embedded edge devices. The proposed method combines the Discrete Wavelet Transform (DWT) with a Wavelet Cross-Correlation (WCC) criterion to extract compact, discriminative features from non-stationary transient pressure signals. Four supervised classifiers are systematically evaluated under the joint constraints of predictive accuracy, memory footprint, and inference latency: Logistic Regression, Decision Trees, Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs). On a laboratory dataset of 454 labeled windows, the SVM achieves the highest accuracy among the tested models while offering substantially lower inference time and memory footprint than the ANN, making it the preferred candidate for embedded deployment. The model is quantized to Q15 fixed-point arithmetic and deployed on a Texas Instruments MSP430FR5994 microcontroller, achieving a 51.3% reduction in model size at the cost of a 114.3% increase in inference time. The total end-to-end processing latency of 21.1 ms is consistent with real-time operation at the tested sampling rate. These results suggest that wavelet-based feature engineering and fixed-point quantization can be optimized to enable practical ML-based burst detection on ultra-low-power microcontrollers.

Keywords: Burst detection, Water distribution networks, Discrete wavelet transform, Support vector machines, Edge computing TinyML, Fixed-point arithmetic

I. INTRODUCTION

Real-time anomaly detection on resource-constrained embedded devices has become a central problem in signal processing and edge machine learning, where inference must run locally under strict energy, memory, and latency budgets. Water distribution networks (WDNs) provide a demanding application domain for this class of problems: pipe bursts are sudden structural failures that produce high-amplitude, non-stationary transient pressure waves, typically lasting less than 2 s, whose delayed detection causes substantial economic, service, and environmental damage [1], [2]. This is acute in infrastructure-limited contexts such as Cuba, where safe drinking water is a recognized human right [3], yet over 50% of the water distributed is lost to leaks [4], [5], [6]. Detecting such transients reliably requires high-frequency sensing, feature extraction, and classification within the computational envelope of low-power microcontrollers; background leaks, by contrast, produce persistent low-amplitude signatures that require more sensitive accelerometric sensing [7].

Burst detection frameworks are broadly classified as model-based or data-driven [8], [9]. Model-based methods rely on calibrated hydraulic models and low-frequency measurements (5–15 min), inadequate for transient events, and their performance degrades severely when accurate network models are unavailable, as in developing infrastructure contexts [10], [11]. Data-driven approaches learn directly from measured signals and, combined with high-frequency sensing, enable accurate transient detection without explicit network models [11]. The Discrete Wavelet Transform (DWT) has emerged as the preferred tool for feature extraction in this context owing to its multiresolution capability, which enables simultaneous time-frequency localization of transient singularities [12]. Machine learning (ML) classifiers applied to wavelet-based features have reported detection rates above 95% with false alarm rates below 1% [1], [13], [14–17]. Beyond binary detection, related studies have addressed leak localization and benchmarked classical statistical methods against ML approaches [18], [19].

A critical open challenge is deploying these pipelines at the network edge, where nodes operate under strict energy, memory, and computational constraints, often without a continuous power supply [20]. Recent surveys on TinyML and on-device inference confirm that ML models can now run on microcontrollers with only kilobytes of RAM, while highlighting open trade-offs among accuracy, model size, and latency [21], [22], [23]. However, integrating high-resolution wavelet feature extraction with lightweight embedded classifiers in a co-designed, end-to-end validated system remains an open problem. This work addresses that gap by jointly optimizing wavelet feature engineering, classifier selection, and fixed-point implementation on a TI MSP430FR5994 microcontroller. The main contributions are: (1) a complete, hardware-aware co-design that integrates DWT/WCC feature extraction, classifier selection, and TinyML deployment into a single real-time pipeline validated end-to-end on a commercial ultra-low-power microcontroller; (2) a fixed-point (Q15) feature-extraction and SVM inference scheme that exploits the MSP430 Low Energy Accelerator and operates within tight memory and execution-time budgets; and (3) a systematic evaluation showing that the proposed SVM attains a 51.3% reduction in model size with no loss of classification performance under joint accuracy and embedded-efficiency constraints.

II. METHODOLOGY

The proposed framework provides an end-to-end pipeline integrating experimental signal acquisition, wavelet-based feature extraction, ML classification, and embedded deployment. DSP stages and ML models are jointly optimized under hardware constraints (co-design), ensuring efficient execution on the target microcontroller.

1. Dataset Description

Pressure signals were acquired on the hydraulic test bench of the Hydraulic Research Center at CUJAE. The setup consists of a high-density polyethylene (HDPE) pipeline (nominal diameter 160 mm, total length 65.2 m) fed by a constant-head tank that maintains a 10 m water column at the inlet. A regulable outlet valve sets a mean

flow velocity of 2 m/s, generating high hydraulic noise and a challenging detection scenario. Two datasets were collected under different sensing configurations, as shown in Figure 1.

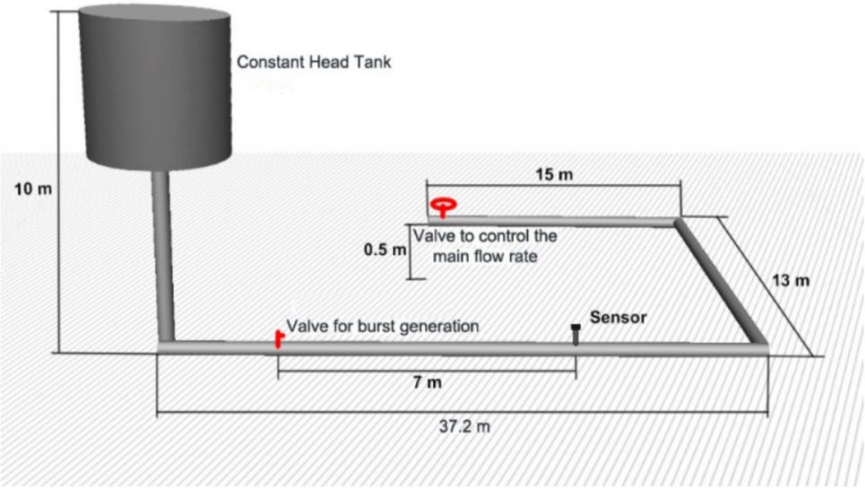


Figure 1. Schematic representation of the experimental setup used for pressure data acquisition.

- Group 1 signals. A Raspberry Pi 3 connected to a MicroSensor MPM3801A digital pressure transducer was used for acquisition at 200 Hz, following the recommendations for ML-based burst detection [13]. Burst events were artificially generated at four distances from the sensor, organized into four subgroups; each of the 120 recordings spans 10 s (2000 samples) and contains a distinct burst event.
- Group 2 signals. An MSP430 microcontroller replaced the Raspberry Pi while retaining the same MPM3801A sensor, positioned 7 m upstream of the simulated leak point. Three leak conditions were defined by flow rates of 0.5, 0.8, and 1.5 L/s; signals were sampled at 400 Hz, and each recording spanned 5 s (2000 samples), yielding 107 files. Combined with Group 1, the full experimental corpus comprises 227 burst recordings.
- Final dataset. Each of the 227 recordings yields one 2-second burst window (Class 1) and one 2-second steady-state noise window (Class 0), producing a balanced dataset of 454 labeled feature vectors, 227 per class. The balanced class distribution eliminates the need for resampling strategies.

2. Data Preprocessing

Preprocessing conditions the raw pressure signals: Group 1 signals were segmented into non-overlapping 2-second windows, sufficient to capture a complete burst transient, and mean-centered to remove the static pressure offset and isolate the dynamic component:

$$x_{centered}[n] = x[n] - \mu_x\,, \quad \mu_x = \frac{1}{N}\sum_1^N x[n] \tag{1}$$

Group 2 signals, originally sampled at 400 Hz, were first decimated to 200 Hz using a 2:1 anti-aliasing filter in MATLAB. This operation halves computational complexity while preserving burst-related content, as transient singularities concentrate their energy well below 100 Hz; the same 2-second windowing and mean-centering was then applied to both groups.

3. Feature Extraction

The DWT was selected for its multiresolution capability, which characterizes local signal singularities at multiple scales simultaneously. Unlike the FFT, which trades away time localization, the DWT preserves both time and frequency resolution [24], essential for non-stationary transients, a choice consistently supported by the burst-detection literature [1], [13], [25], [26]. The DWT is efficiently computed through the Mallat filter bank algorithm [13], whose successive low-pass/high-pass filtering and 2:1 downsampling stages can be accelerated by the Low Energy Accelerator (LEA) of the MSP430FR5994, offloading the CPU from computationally intensive convolutions. A comparative analysis confirmed that for signal lengths between 32 and 256 samples, time-domain convolution outperforms FFT-based methods by 65.3% on the LEA module, thereby motivating a direct filter-bank implementation. The Mallat scheme is illustrated in Fig. 2.

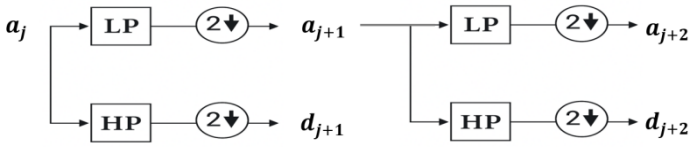


Figure 2. Mallat scheme for computing the DWT [13].

1) Feature extraction strategy. The algorithm computes detail coefficients at decomposition levels 4 and 5 (d_4, d_5), corresponding to the frequency bands [12.5, 25] Hz and [6.25, 12.5] Hz, respectively, at a sampling rate of 200 Hz, where burst-induced singularities concentrate their energy. The WCC is then computed between d_4 and d_5 after linear-interpolation alignment to equalize their lengths, and the modulus $|WCC_4|$ serves as the feature descriptor, following [25]. As illustrated in Figure 3, the burst around the 5th second coincides with a clear $|WCC_4|$ peak, confirming the discriminative power of this descriptor. The wavelet basis was selected through Lipschitz exponent analysis [27]: among multiple candidate wavelets, Coiflet 1 yielded the lowest exponent, indicating stronger singularity enhancement and transient localization.

2) Reference C implementation. A floating-point reference implementation was first developed in standard C on a workstation to establish a numerical baseline free from embedded constraints, to process all 454 windows into a standardized CSV dataset (30 $|WCC_4|$ features plus a binary label per row), and to provide a portable codebase for the embedded target. The core uses a cascaded filter bank with buffer reuse across levels, with WCC computed via interpolation alignment, point-wise multiplication, and magnitude extraction.

3) MSP430FR5994 embedded implementation. The embedded adaptation introduces Q15 fixed-point arithmetic: a 16-bit signed representation (1 sign bit, 15 fractional bits) covering $[-1, \approx 0.99997]$; conversion from floating-point involves normalization to $[-1, 1]$, saturation clamping, and scaling by 32767 with rounding. The MSP-DSPLIB functions msp_fir_q15(), msp_mpy_q15(), msp_abs_q15(), and msp_deinterleave_q15() provide LEA-accelerated fixed-point operations; during LEA execution, the CPU sleeps in LPM0, waking only on interrupt, minimizing energy consumption. All SVM parameters are stored in persistent FRAM sections via #pragma PERSISTENT, enabling efficient access during repeated kernel evaluations. As shown in Figure 4, the embedded implementation exhibits higher $|WCC_4|$ amplitudes near signal singularities compared to the floating-point reference, attributed to improved convolution edge handling by msp_fir_q15(), which yields more discriminative feature vectors.

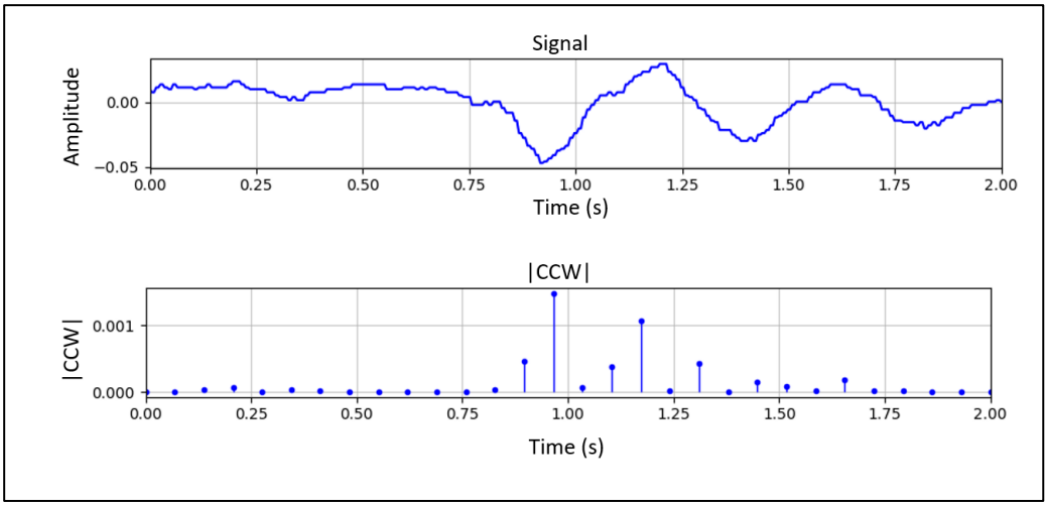


Figure 3. Feature extraction of a pressure signal: raw pressure signal containing a burst event around t = 5 s (top) and the corresponding |WCC$_4$| modulus (bottom), whose dominant peak coincides in time with the burst transient.

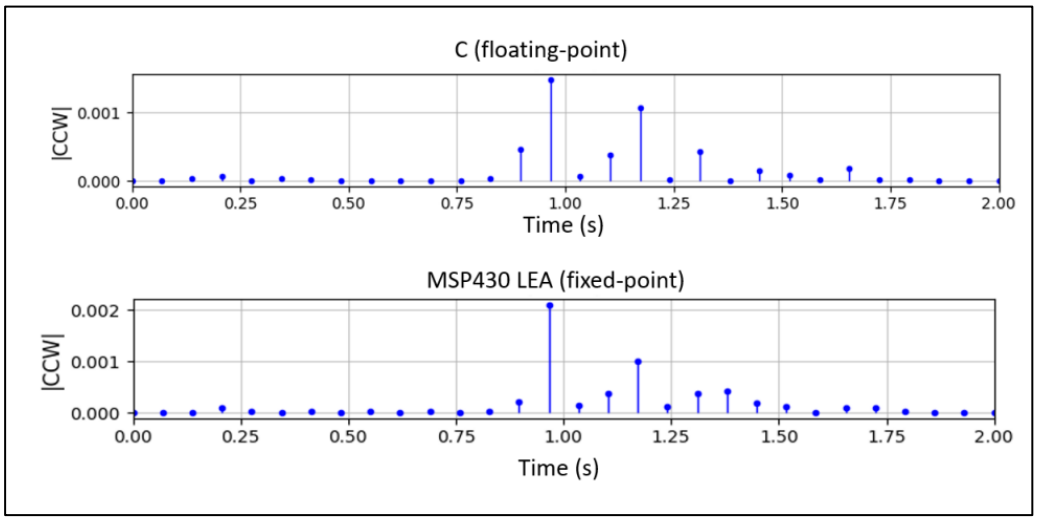


Figure 4. Computation of |WCC$_4$| of the same burst signal: floating-point workstation reference (top) versus Q15 fixed-point implementation on the MSP430 LEA (bottom); both implementations localize the burst at the same coefficient positions.

4. Classification Framework

The 30-dimensional |WCC$_4$| feature vectors trained four classifiers implemented in scikit-learn within a Google Colab environment. To rule out train–test contamination, the dataset was split once into stratified training and test subsets (85%/15% [28]), preserving class balance; the resulting 68-sample test set remained completely unseen during model design. Hyperparameter optimization was confined to the training subset, using exhaustive GridSearchCV with stratified k-fold cross-validation [29], which averages performance across k train/validation partitions and is particularly valuable for medium-sized datasets. Only after this selection was finalized were the models evaluated once on the untouched test set, so the reported metrics reflect generalization to unseen data.

Logistic Regression: configured as a linear classifier with ℓ2 regularization ($C = 10.0$, liblinear solver), penalizing large coefficients to reduce overfitting.

Decision Tree: employs the Gini impurity criterion with a maximum depth of 8 levels; pre-pruning requires at least 5 samples per split and 2 per leaf, balancing expressiveness and generalization.

Artificial Neural Network: a sequential architecture with two hidden layers (64 and 32 neurons), dropout regularization of 30% and 20%, ReLU activations, Adam optimizer, and binary cross-entropy loss; early stopping (patience 10 epochs) monitors validation loss to prevent overfitting. Due to the high-dimensional hyperparameter search space, manual tuning was used instead of GridSearchCV.

Support Vector Machine: implemented via scikit-learn's SVC with an RBF kernel ($C = 10$, $\gamma = 1$) optimized by GridSearchCV. The trained model relies on 152 support vectors, limiting storage to only these vectors, their dual coefficients, and the bias term.

Embedded SVM on MSP430FR5994. Following evaluation, the SVM was ported to the MSP430FR5994 by exporting model parameters from scikit-learn: 152 support vectors (each 30-dimensional), dual coefficients, a bias term, and RBF kernel parameters, which were stored as constant Q15 arrays in FRAM. Inference evaluates the RBF kernel against each support vector and computes the weighted sum plus bias, entirely without external library dependencies.

III. RESULTS

1. Classifier Performance

Table 1 reports test metrics for all four classifiers. Logistic Regression reaches 98% accuracy (one misclassified sample), while the Decision Tree shows the lowest performance (three false positives, one false negative), consistent with the limited generalization of greedy splitting. Both the ANN and the SVM classify every test sample correctly, achieving 100% across all metrics. This confirms that the 30-dimensional $|WCC_4|$ space is sufficiently discriminative to support perfect separation between burst events and background noise.

Table 1. Performance metrics on the independent test dataset

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Logistic Regression	98%	100%	97%	98%	99%
Decision Tree	94%	93%	97%	94%	93%
ANN	100%	100%	100%	100%	100%
SVM	100%	100%	100%	100%	100%

Table 2 presents the absolute differences between cross-validation and test performance, measuring model stability across splits. Logistic Regression and Decision Tree exhibit discrepancies of up to 10% in recall, indicating sensitivity to the particular training-test partition. ANN and SVM show minimal variation (no more than 2% in accuracy, 0% in recall), confirming that their perfect test performance reflects genuine generalization rather than favorable data splitting.

Table 2. Absolute differences: cross-validation minus test metrics

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Logistic Regression	5%	5%	6%	5%	3%
Decision Tree	7%	6%	10%	7%	5%

ANN	2%	6%	0%	3%	2%
SVM	2%	6%	0%	3%	2%

Given the equivalent predictive performance of ANN and SVM, model selection was based on operational efficiency: the ANN requires 82.1 KB and 132.2 ms inference time, whereas the SVM requires only 41.3 KB and 6.16 ms, roughly 21× faster and 50% smaller. The SVM was therefore selected for embedded deployment, providing the best trade-off between classification performance, memory footprint, and computational efficiency [30]. The latency gap follows from model structure: ANN inference propagates the input through two fully connected hidden layers, whereas the SVM reduces inference to 152 RBF kernel evaluations followed by a single weighted sum, a far smaller and more regular workload on the microcontroller.

2. Embedded Validation

The same 68 test signals were processed end-to-end on the MSP430FR5994: each raw pressure window was passed through the embedded Q15 feature extraction pipeline and classified by the embedded SVM. The confusion matrix (Table 3) shows perfect classification of all 34 normal and 34 burst samples, confirming that the pipeline preserves its predictive integrity under fixed-point representation and embedded constraints.

Table 3. Confusion matrix of the SVM on MSP430FR5994

	Predicted: Normal	Predicted: Burst
Actual: Normal	34	0
Actual: Burst	0	34

3. Cloud vs. Embedded Performance

Table 4 compares the floating-point cloud implementation against the Q15 embedded implementation. Quantization achieves a 51.3% reduction in model size (41.3 KB to 20.1 KB), enabling the model to fit entirely within the 256 KB FRAM of the target device while maintaining a 100% detection rate and 0% false-positive rate. Inference time increases from 6.16 ms to 13.2 ms (+114.3%), attributable to the per-multiply scaling, rounding, and saturation overhead of fixed-point kernel evaluation, which cannot yet fully exploit LEA parallelism; the size reduction, in turn, follows from 16-bit parameters halving the storage of support vectors and coefficients.

Table 4. Cloud (floating-point) vs. embedded (Q15) comparison

Metric	Cloud	Embedded	Difference
Detection Rate	100%	100%	0%
Model Size	41.3 KB	20.1 KB	−51.3%
Inference Time	6.16 ms	13.2 ms	+114.3%

4. Temporal Processing Distribution

Table 5 details the execution time breakdown of the complete on-device pipeline. The total latency of 21.1 ms allows each 2-second window to be processed and classified well within real-time constraints for 200 Hz monitoring. SVM inference is the most demanding stage at 13.2 ms (62.5%), driven by 152 RBF kernel evaluations over 30-dimensional vectors. Preprocessing requires 7.9 ms in total, with approximation coefficient extraction dominating at 5.14 ms (24.3%) due to cascaded filter-bank convolutions at levels 1–4. Interpolation, WCC multiplication, and magnitude computation add 1.80 ms (8.5%). Feature extraction and SVM inference

together account for over 86% of the total latency, making the approximation-coefficient filter bank and the RBF kernel evaluation the natural targets for further LEA-based acceleration.

Table 5. Execution time distribution on MSP430FR5994

Operation	Time (ms)	Percentage
Approximation coefficients (levels 1–4)	5.14	24.3%
Detail coefficients (d_4, d_5)	0.99	4.7%
Interpolation + WCC + magnitude	1.80	8.5%
SVM inference (152 RBF evaluations)	13.2	62.5%
Total	21.1	100%

5. Memory Utilization

Table 6 reports memory utilization on the MSP430FR5994. FRAM is only 10.9% utilized, primarily for the SVM parameters stored via #pragma PERSISTENT. In comparison, SRAM is 43.3% utilized, mainly by the feature-extraction buffers (filter coefficients, intermediate decomposition arrays, and the output feature vector). The remaining headroom leaves substantial capacity for communication stacks, event logging, or algorithm extensions.

Table 6. Memory utilization on MSP430FR5994

Memory	Total	Used	Available
FRAM	256 KB	27.48 KB (10.9%)	228.52 KB (89.1%)
SRAM	8 KB	3.47 KB (43.3%)	4.53 KB (56.7%)

Discussion. The results show that a complete burst-detection pipeline (wavelet-based feature extraction, classifier selection, and inference) can run entirely on an ultra-low-power microcontroller while preserving the predictive performance of its floating-point cloud counterpart. This is the central strength of the approach: detection accuracy and embedded feasibility are not competing objectives, since fixed-point quantization compresses the model without degrading classification, and inference still completes within the real-time budget. The combination of a compact $|WCC_4|$ descriptor, fixed-point arithmetic, and LEA-accelerated processing thus offers a practical basis for autonomous, in-network monitoring, with ample memory headroom for communication and logging. These findings should be read within the scope of the study: the experiments used a single laboratory test bench with a controlled HDPE pipeline and artificially induced bursts, whereas operational networks differ in pipe materials, pressures, and background noise. The perfect classification reported here, therefore, establishes feasibility on the target hardware rather than field performance: the held-out test set contains 68 windows, so the reported metrics carry the statistical uncertainty of a small sample and were not accompanied by a significance analysis across repeated data partitions. This reading is consistent with related embedded ML studies on water infrastructure, which report high but imperfect accuracies on more heterogeneous data [31], [32], and with field deployments of data-driven burst detection in operational networks, where performance is assessed against repair records and a considerable share of alerts does not correspond to confirmed bursts [33]. Validation on operational networks with a larger, more varied dataset, together with statistical robustness analysis across repeated splits and benchmarking against additional lightweight classifiers such as gradient-boosted trees and compact one-dimensional convolutional networks, is the natural next step, made feasible by the solution's low cost and power consumption.

IV. CONCLUSION

This work has presented a co-designed framework for real-time burst detection in WDNs and has been validated end-to-end on an MSP430FR5994 ultra-low-power microcontroller. The combination of DWT and WCC applied to detail coefficients at levels 4 and 5, using a Coiflet-1 basis, yields a compact 30-dimensional feature descriptor that achieves clear separation between burst events and background noise, with the targeted frequency sub-bands (6.25–25 Hz) physically aligned with the energy concentration of burst-induced transients.

Among the four evaluated classifiers, ANN and SVM achieve 100% accuracy on the independent test dataset with minimal cross-validation variability. The SVM was selected for embedded deployment owing to its 21× lower inference latency and 50% smaller footprint, with no sacrifice in predictive performance. Q15 quantization further reduces the embedded model size by 51.3% with no classification performance loss, confirming precision reduction as a viable compression strategy for TinyML deployment. The end-to-end latency of 21.1 ms remains well within real-time constraints for 200 Hz monitoring, and the pipeline occupies only 10.9% of FRAM and 43.3% of SRAM, leaving substantial headroom for communication or logging modules.

Future work will address multi-sensor network topologies for burst localization, further latency reduction through LEA-optimized RBF kernel computation, online adaptation strategies to handle hydraulic operating condition drift, and field validation on real WDN infrastructure with heterogeneous pipe materials and pressures.

REFERENCES

[1]. J. R. Beltrán et al., “Detección, Identificación y Localización de Eventos en Tuberías de Agua con Alta Densidad Temporal,” in Proc. Congreso Internacional de Ingeniería Eléctrica y Electrónica (ELECTRO), La Habana, Cuba, oct. 2017.

[2]. F. Boztáş et al., “Analyzing the effect of the unreported leakages in service connections of water distribution networks on non-revenue water,” Int. J. Environ. Sci. Technol., vol. 16, pp. 4393–4406, 2019.

[3]. United Nations General Assembly, “The human right to water and sanitation,” Resolution 64/292, United Nations, New York, Jul. 2010.

[4]. D. Duque y J. Antonio, “El agua en Cuba: un desafío a la sostenibilidad,” Ingeniería Hidráulica y Ambiental, vol. 39, no. 2, pp. 46–59, 2018.

[5]. R. S. Rivas, “La capital provincial con los ciclos de agua más largos del país”, Granma.cu, 2021.

[6]. Cuál es la situación del abasto de agua en Cuba?” Cubadebate, Apr. 2024.

[7]. C. A. Segarra Cabrera, “Diseño de un sistema de detección de fugas mediante valores frecuenciales,” M.Sc. thesis, UPS, 2024.

[8]. Y. Wu and S. Liu, “A review of data-driven approaches for burst detection in WDS,” Urban Water J., vol. 14, no. 9, pp. 972–983, 2017, doi: 10.1080/1573062X.2017.1279191.

[9]. W. Nimri et al., “Data-driven approaches for leak detection in WDS: a review,” Neural Comput. Appl., vol. 35, pp. 11611–11623, 2023, doi: 10.1007/s00521-023-08497-x.

[10]. D. Jung and K. Lansey, “Water distribution system burst detection using a nonlinear Kalman filter,” J. Water Resour. Plan. Manag., vol. 141, no. 5, art. 04014070, 2015, doi: 10.1061/(ASCE)WR.1943-5452.0000464.

[11]. Z. Hu et al., “Review of model-based and data-driven approaches for leak detection,” Water Supply, vol. 21, no. 7, pp. 3282–3306, 2021, doi: 10.2166/ws. 2021.101.

[12]. S. Srirangarajan et al., “Wavelet-based burst event detection in WDS,” J. Signal Process. Syst., vol. 72, no. 1, pp. 1–16, 2013, doi: 10.1007/s11265-012-0690-6.

[13]. I. Vega González et al., “Algorithm to detect bursts in water pipes for low-power devices,” Rev. Ing. Electrón. Automát. Comun. (RIELAC), vol. 43, no. 1, art. e2204, 2022.

[14]. T. Ravichandran et al., “Ensemble-based ML for improved leak detection,” J. Hydroinform., vol. 23, no. 2, pp. 307–323, 2021.

[15]. P. Huang et al., "Real-time burst detection in DMAs using ML," Water, vol. 10, no. 12, art. 1765, 2018.
[16]. M. Quiñones-Grueiro et al., "An unsupervised approach to leak detection in WDNs," Int. J. Appl. Math. Comput. Sci., vol. 28, no. 2, pp. 283–295, 2018.
[17]. V. Muralidharan and V. Sugumaran, "Feature extraction using wavelets and classification through decision tree algorithm for fault diagnosis of mono-block centrifugal pump," Measurement, vol. 46, no. 1, pp. 353–359, 2013, doi: 10.1016/j.measurement.2012.07.007.
[18]. A. Soldevila et al., "Data-driven approach for leak localization in WDNs," Water, vol. 11, no. 7, art. 1500, 2019, doi: 10.3390/w11071500.
[19]. J. C. Van der Walt et al., "Pipe network leak detection: statistical vs. ML," Urban Water J., vol. 15, no. 10, pp. 953–960, 2018, doi: 10.1080/1573062X.2019.1597375.
[20]. M. Allen et al., "Real-time in-network distribution system monitoring," J. AWWA, vol. 103, no. 7, pp. 63–75, 2011.
[21]. N. N. Alajlan and D. M. Ibrahim, "TinyML: enabling inference on ultra-low-power IoT devices," Micromachines, vol. 13, 2022.
[22]. S. Heydari and Q. H. Mahmoud, "Tiny Machine Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions," Sensors, vol. 25, no. 10, art. 3191, 2025.
[23]. A. Elhanashi, P. Dini, S. Saponara, and Q. Zheng, "Advancements in TinyML: Applications, Limitations, and Impact on IoT Devices," Electronics, vol. 13, no. 17, art. 3562, 2024.
[24]. S. Zhou et al., "Comparison between FFT and wavelet analysis for non-stationary signals," in Proc. ASIA 2009.
[25]. E. Trutié-Carrero et al., "Detección y Localización de Ruptura Súbita mediante DWT y Correlación Cruzada," RIAI, vol. 15, 2018.
[26]. U. Rajasekaran and M. Kothandaraman, "A survey and study of signal and data-driven approaches for pipeline leak detection and localization," J. Pipeline Syst. Eng. Pract., vol. 15, no. 2, art. 03124001, 2024.
[27]. E. Trutié-Carrero et al., "Automatic burst detection in WDS by Lipschitz exponent and WCC," Measurement, vol. 151, art. 107195, 2020, doi: 10.1016/j.measurement.2019.107195.
[28]. C. Miller et al., "A review of model evaluation metrics for ML in genomics," Front—Bioinform., vol. 4, 2024.
[29]. L. Sweet et al., "Cross-validation strategy impacts ML model performance," Artif. Intell. Earth Syst., vol. 2, 2023.
[30]. S. Suthaharan, "Support vector machine," in Machine Learning Models for Big Data Classification, Springer, 2016.
[31]. O. Atanane, A. Mourhir, N. Benamar, and M. Zennaro, "Smart buildings: water leakage detection using TinyML," Sensors, vol. 23, no. 22, art. 9210, 2023, doi: 10.3390/s23229210.
[32]. J. Alves Coelho, A. Glória, and P. Sebastião, "Precise water leak detection using machine learning and real-time sensor data," IoT, vol. 1, no. 2, pp. 474–493, 2020, doi: 10.3390/iot1020026.
[33]. S. R. Mounce, J. B. Boxall, and J. Machell, "Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows," J. Water Resour. Plan. Manag., vol. 136, no. 3, pp. 309–318, 2010, doi: 10.1061/(ASCE)WR.1943-5452.0000030.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Role-Based Phishing to Ransomware Readiness for Clinical Staff

Leon Carpelan
Laurea University of Applied Sciences
Espoo, Finland
Leon.Carpelan@student.laurea.fi

Anderson Siqueira Teixeira
Laurea University of Applied Sciences
Espoo, Finland
Anderson.Siqueira.Teixeira@student.laurea.fi

Wimpie Roothe
Laurea University of Applied Sciences
Espoo, Finland
Wimpie.Roothe@student.laurea.fi

Robert Simonsuuri
Laurea University of Applied Sciences
Espoo, Finland
Robert.Simonsuuri@student.laurea.fi

Jyri Rajamäki
Laurea University of Applied Sciences
Espoo, Finland
Jyri.Rajamaki@laurea.fi

Abstract
Hospitals are common targets for cybercriminals, and most ransomware attacks in healthcare start with a phishing email. Phishing (sending fake emails or messages designed to trick users into clicking malicious links, opening infected attachments, or revealing sensitive information such as login credentials) can give criminals access to hospital systems or allow malicious software, including ransomware, to spread through the network and disrupt operations. The problem is that most cybersecurity training programs are for the IT specialists, not for nurses or doctors. International guidelines such as the NIST Cybersecurity Framework and ENISA guidance are useful at an organizational level, but do not tell a ward nurse what to do when she receives a suspicious email. This paper describes a Work-in-Progress model called the Role-Based Phishing-to-Ransomware Readiness Model, built around a ward nurse persona developed within the XiA (Xpanding Innovative Alliance) project. This XiA approach focuses on developing human-centric, explainable AI systems that support decision-making in complex operational environments. The XiA persona (a typical ward nurse supported by an AI-based assistant that provides clear, context-aware, and explainable guidance in everyday clinical situations) is used to translate abstract cybersecurity concepts into practical, role-specific actions that are understandable and relevant for clinical staff without technical expertise. The developed model combines a visual attack chain, a role-based action table, and a short training scenario to help clinical staff recognize phishing attempts and respond appropriately in real-world situations. Testing with real staff is planned as future work.

Keywords: phishing, ransomware, healthcare cybersecurity, clinical staff, role-based training, XiA persona

I. INTRODUCTION

Hospitals increasingly rely on interconnected digital systems to support patient care, diagnosis, communication, and administrative processes. While this digital transformation improves the efficiency and quality of healthcare services, it also expands the attack surface available to cybercriminals. Among the most common cyber threats targeting healthcare organizations are phishing attacks, which attempt to trick users into revealing credentials, opening malicious attachments, or clicking harmful links. Once access is obtained, attackers may deploy ransomware that disrupts clinical operations and jeopardizes patient safety. A recent example occurred in London in 2024, where a ransomware-related cyberattack led to the cancellation of more than 800 planned medical procedures within a single week [1]. Previous research has similarly demonstrated that human behavior

often determines whether an initial phishing incident develops into a large-scale ransomware event [2]. As healthcare organizations continue to strengthen their technical cybersecurity capabilities, increasing attention has been paid to the human dimension of cyber resilience. Existing cybersecurity frameworks, including the NIST Cybersecurity Framework [3] and ENISA guidance [4], provide valuable organizational and governance-level recommendations. However, these frameworks typically focus on risk management, policies, and technical controls rather than on the practical decisions that frontline healthcare workers must make when confronted with suspicious emails or other potential cyber threats during routine clinical work. Consequently, nurses, physicians, and other clinical personnel often receive cybersecurity training that is either overly generic or overly technical, undermining real-time decision-making in healthcare environments characterized by time pressure, interruptions, and high workloads. Although phishing awareness, ransomware preparedness, and cybersecurity education have been widely studied, an important gap remains between organizational cybersecurity guidance and role-specific operational support for frontline healthcare professionals. Existing studies commonly emphasize awareness training, organizational policies, or general user behavior. In contrast, relatively little attention has been given to translating phishing-to-ransomware attack pathways into practical, role-based guidance tailored to specific healthcare professions. Furthermore, the potential of persona-based educational approaches and explainable learning mechanisms for supporting cybersecurity readiness among clinical staff remains underexplored. This study addresses this gap by developing a Role-Based Phishing-to-Ransomware Readiness Model for clinical staff. The work is connected to the XiA (Xpanding Innovative Alliance) project. This European initiative promotes human-centered and explainable approaches for strengthening competencies related to the European Health Data Space (EHDS) [4]. Within the XiA framework, personas are used to bridge the gap between complex technical concepts and the practical needs of different user groups. In this study, a ward nurse persona is employed to represent a typical frontline healthcare professional who may encounter phishing attacks during everyday clinical work. The proposed model combines a healthcare-specific phishing-to-ransomware attack chain, a persona-based learning perspective inspired by XiA principles, and a role-based action framework designed to support practical decision-making during cyber incidents. Rather than focusing on technical countermeasures, the model seeks to translate cybersecurity knowledge into clear, actionable guidance for non-technical clinical personnel.

The contributions of this Work-in-Progress paper are fourfold:

- A healthcare-specific phishing-to-ransomware readiness model that visualizes how phishing incidents may escalate into ransomware attacks.
- A XiA-inspired persona-based approach for translating cybersecurity concepts into practical guidance for frontline clinical staff.
- A role-based action framework that links different stages of an attack to concrete response actions understandable by non-technical healthcare professionals.
- A foundation for future XiA-enabled intelligent assistants capable of providing explainable and context-aware cybersecurity support within healthcare environments.

The remainder of this paper is organized as follows. Section II reviews the relevant literature on phishing, ransomware, healthcare cybersecurity training, and the XiA project. Section III presents the Design Science Research approach used to develop the proposed model. Section IV describes the resulting artifacts, including the attack chain, persona-based learning framework, role-based action table, and training scenario. Section V discusses the current status of the work, limitations, and future research directions. Finally, Section VI concludes the paper and outlines opportunities for empirical validation and further integration within the XiA framework.

II. BACKGROUND AND RELATED WORK

Healthcare is one of the primary targets of cybercrime because hospitals store sensitive personal data and depend on continuous operations for patient care. As a result, healthcare organizations are often considered more likely

to pay a ransom following a cyberattack, making them attractive targets for ransomware groups [5]. Phishing remains the most common initial attack vector through which ransomware gains access to healthcare systems [2].

Research has consistently shown that human behavior plays a critical role in cybersecurity outcomes. Despite this recognition, most existing frameworks and training materials are designed for IT professionals and focus on organizational controls rather than everyday decision-making in clinical work [6]. Prümmer, van Steen, and van den Berg [7] conducted a meta-analysis of cybersecurity training studies. They found that traditional training programs typically improve users' knowledge but do not reliably change their behavior in stressful real-world situations. In healthcare environments characterized by time pressure, frequent interruptions, and high workloads, this gap between knowing and acting can significantly increase cybersecurity risks. Recent research has also examined the effectiveness of organizational phishing-awareness training in real operational environments. Hillman, Harel, and Toch [8] conducted a controlled experiment involving approximately 5,000 employees in a large financial institution and found that users were significantly more likely to engage with phishing emails containing personalized content. Their results further showed that phishing susceptibility varied across organizational units, while the timing of awareness training alone did not significantly affect phishing click-through rates. The study emphasizes the importance of data-driven, role-sensitive awareness programs. It suggests that employee behavior remains a critical factor in organizational cyber resilience despite advances in technical security controls. These findings further support the need for tailored and context-aware cybersecurity guidance rather than generic awareness training. The importance of actionable guidance is also supported by research grounded in Protection Motivation Theory (PMT). Van Bavel et al. [9] studied the effects of behavioral nudges on online security behavior using a large-scale experiment involving more than 2,000 participants from five European countries. Their findings showed that messages that provided clear coping strategies were significantly more effective at promoting secure behavior than messages that focused primarily on threat consequences. Furthermore, coping-focused messages were found to be as effective as combining both threat and coping messages. These results suggest that cybersecurity interventions should not only raise awareness of risks but also provide users with simple and actionable recommendations. This finding aligns with the objective of the present study, which seeks to translate cybersecurity threats into role-based actions that healthcare professionals can apply in everyday clinical situations. Both the NIST Cybersecurity Framework [3] and ENISA guidance [4] acknowledge the importance of the human factor in cybersecurity. However, these frameworks primarily focus on governance, risk management, and technical controls rather than providing practical instructions for frontline healthcare workers. Similarly, Syafrizal et al. [10] observed that many cybersecurity standards inadequately address role-specific human vulnerabilities. Kruse et al. [11] further found that healthcare professionals benefit most from training that is directly connected to their daily work practices rather than from generic awareness programs. Taken together, these findings suggest that effective cybersecurity education should provide practical, role-oriented guidance tailored to the responsibilities and decision-making contexts of specific user groups. This role-based perspective aligns closely with the objectives of the XiA (Xpanding Innovative Alliance) Project [12]. XiA recognizes that interoperability is not solely a technical challenge involving standards and technologies, but also a human challenge requiring professionals to understand, apply, and further develop interoperability concepts in real-world settings. To address this need, the project combines EHDS-aligned interoperability expertise with innovative educational approaches, including modular learning building blocks, flexible micro-learning pathways, standards-based practical content, and progressive recognition of competencies through micro-credentials. These principles are intended to support professionals, educators, and organizations in developing and maintaining competencies over time. A key principle of XiA is that digital transformation requires different stakeholder groups to acquire different competencies [13]. Consequently, XiA categorizes learners into three broad groups: developers, implementers, and end-users. Developers focus on building and maintaining interoperable digital health systems, implementers are responsible for planning, integrating, and governing such systems, and end-users include healthcare professionals who use digital solutions in their daily clinical practice. This learner-centered approach supports the development of personalized learning pathways tailored to specific professional roles and responsibilities [13]. The present study adopts this XiA-inspired perspective by focusing on an end-user persona representing a ward nurse. Rather than

presenting cybersecurity concepts through technical terminology or organizational policies, the study translates phishing and ransomware threats into practical, role-specific guidance that reflects the realities of clinical work. In doing so, the proposed approach seeks to bridge the gap between cybersecurity knowledge and everyday decision-making, thereby supporting safer, more informed behavior among frontline healthcare professionals.

III. SOLUTION AND APPROACH

This study adopted a Design Science Research (DSR) approach [14] to develop a role-based educational model for improving phishing-to-ransomware readiness among clinical staff, as illustrated in Figure 1. The model was developed through iterative interaction between healthcare cybersecurity requirements, XiA-based educational principles, and established cybersecurity knowledge. The building process consisted of four phases. First, a literature review was conducted to examine academic research and authoritative guidance from ENISA and NIST. The objective was to identify how phishing attacks evolve into ransomware incidents in healthcare environments and to determine the limitations of existing cybersecurity training approaches. Second, the authors further developed a XiA-based persona representing a typical ward nurse. Personas are fictional yet realistic representations of target users that help translate abstract concepts into practical, relatable scenarios. The selected persona reflects the daily reality of many healthcare professionals: working under time pressure, using shared clinical workstations, and frequently receiving emails related to patient information, laboratory results, or shift schedules. These characteristics also correspond to common phishing lures used by cybercriminals. The persona attributes were derived from findings reported in healthcare cybersecurity literature, particularly studies describing clinical workflows, workload pressures, email usage patterns, and common phishing exposure factors among nursing staff.

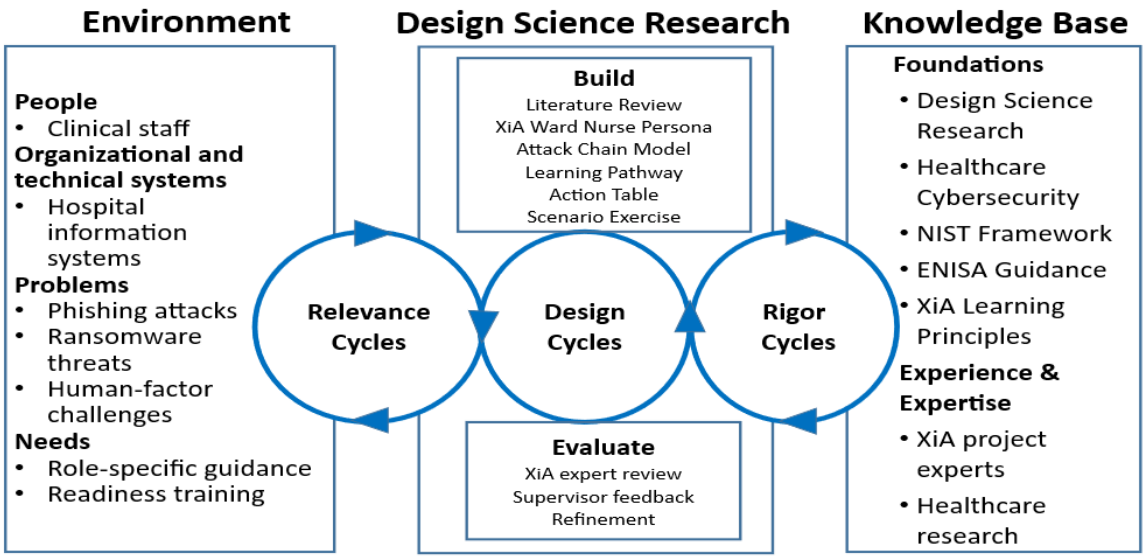


Figure 1. Design science research process applied to the development of the role-based phishing-to-ransomware readiness model.

Third, the findings from the literature review and the ward nurse persona were combined to create a visual phishing-to-ransomware attack chain, a XiA-enabled role-based cybersecurity learning framework, and a role-based action table. These artifacts were designed to transform cybersecurity knowledge into clear, actionable guidance adapted to the needs and responsibilities of frontline clinical staff. Finally, the proposed model was reviewed by supervisors and experts associated with the XiA project. Their feedback was used to refine the

model and improve its clarity, usability, and alignment with the XiA principles. The main outputs of this Work-in-Progress study are the six-stage phishing-to-ransomware attack chain presented in Figure 2, the XiA-enabled role-based cybersecurity guidance framework presented in Figure 3, the role-based action table presented in Table 2, and the training scenario described in Section IV.5.

IV. RESULTS

This section presents the key components of the developed model, explaining how phishing attacks can evolve into ransomware incidents in a healthcare setting and how clinical staff can intervene at different stages. The section combines attack pathways identified in the literature with a user-centered perspective informed by the XiA persona and practical response guidance. The goal is to translate abstract cybersecurity knowledge into clear, role-specific actions that can be directly applied in everyday clinical work.

1. The Attack Chain

Figure 2 presents the six-stage phishing-to-ransomware attack chain identified from the literature. The model illustrates how a phishing email can lead to credential compromise, unauthorized system access, ransomware deployment, and ultimately disruption of healthcare operations. The objective is to support clinical staff in recognizing and potentially interrupting the attack during its early stages.

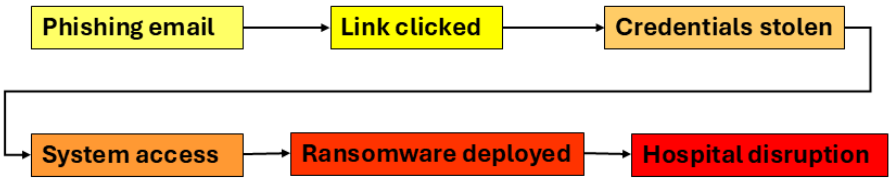


Figure 2. Six-stage phishing-to-ransomware attack chain in a healthcare setting.

2. Mapping to Established Cybersecurity Frameworks

To strengthen the theoretical grounding of the proposed model, the six-stage phishing-to-ransomware attack chain was mapped against two widely recognized cybersecurity frameworks: the Lockheed Martin Cyber Kill Chain [15] and the MITRE ATT&CK framework [16]. The Cyber Kill Chain describes the progression of a cyberattack from initial delivery to the achievement of the attacker's objectives. At the same time, MITRE ATT&CK provides a knowledge base of adversarial tactics and techniques observed in real-world cyber incidents. The comparison demonstrates that the proposed attack chain does not introduce a new technical taxonomy of cyberattacks. Instead, it adapts established attack concepts into a healthcare-oriented educational artifact designed to support frontline clinical staff. By translating technical attack stages into role-specific, actionable concepts, the model aims to improve the clarity and practical relevance of cybersecurity guidance for non-technical healthcare professionals.

Table 1 presents the mapping between the proposed attack chain and the corresponding phases of the Cyber Kill Chain and MITRE ATT&CK frameworks.

Table 1. Mapping of the proposed attack chain to established cybersecurity frameworks

Proposed Stage	Cyber Kill Chain [15]	MITRE ATT&CK [16]
Phishing Email	Delivery	Initial Access
Link Clicked	Exploitation	Execution
Credentials Stolen	Exploitation / Credential Theft	Credential Access
System Access	Actions on Objectives	Persistence / Lateral Movement
Ransomware Deployed	Actions on Objectives	Impact
Hospital Disruption	Actions on Objectives	Impact

The mapping confirms that each stage in the proposed model corresponds to attack behaviors already recognized in established cybersecurity literature and practice. The novelty of the present work, therefore, lies not in defining new attack stages but in integrating these stages with XiA-inspired personas, role-based guidance, and healthcare-specific learning materials. This approach transforms technical cybersecurity knowledge into a more accessible and actionable format for clinical end-users.

3. Role Exposure and the XiA Persona

Not all hospital staff face the same phishing risk. Table 2 compares phishing exposure across four healthcare roles based on findings from [2] and [17]. The XiA ward nurse persona represents a high-exposure role due to frequent interaction with time-sensitive communications, including Electronic Health Record (EHR) notifications.

Table 2. Comparison of phishing exposure across healthcare roles (adapted from [2] and [17])

Role	Common phishing lure	Exposure level	Response Urgency
Ward nurse	Lab result/rota update	High	Very High
Physician	Referral / EHR notice	High	Very High
Admin staff	Invoice / HR/supplier email	Medium	Moderate

Role	Common phishing lure	Exposure level	Response Urgency
IT support	Password reset/support ticket	Critical	Immediate

4. XiA-enabled Cybersecurity Guidance

While Figure 2 illustrates how a phishing attack can progress to ransomware, Figure 3 demonstrates how XiA principles can be used to provide explainable, context-aware guidance that helps users interrupt the attack chain before it reaches the disruption stage.



Figure 3. XiA-enabled role-based cybersecurity learning framework.

5. Role-based Action Table

Table 3 provides a simple action for each clinical staff member at each stage of an attack. The actions are written in plain language because recent research [7] shows that short, specific instructions work much better in clinical situations than long policy documents. The key message is that reporting a mistake early is much better than staying silent. When a staff member immediately reports a suspicious link to the IT team after clicking it, the security team can start blocking the attack before it spreads. Early reporting is one of the most effective ways to limit damage [4].

Table 3. Role-based response actions for clinical staff (adapted from [17] and [18])

Stage	Situation	Recommended action
Phishing email	You received a suspicious message	Do not click any links or open attachments. Report the email to IT.
Link clicked	You may have clicked something bad.	Disconnect your device from the network right away. Then report to IT immediately.

Stage	Situation	Recommended action
Credentials at risk	Your login details may have been stolen	Change your password now and call IT. Do not wait.
Strange behaviour	Your computer is acting unusually.	Stop using the device, but do not shut it down. Inform IT as soon as possible.
Ransomware	Systems are locked, or files cannot be opened.	Follow your hospital emergency cyber response procedure. Do not try to fix it yourself.

6. Scenario Exercise

A nurse arrives for her shift and checks her email. She notices a message labeled “Urgent: patient test results, action required.” Because the message appears legitimate, and she is working under time pressure, she clicks the link and enters her credentials on a fake login page. A few hours later, hospital systems begin to malfunction, and staff are unable to access patient records due to ransomware. The scenario is used alongside Table 3 to help participants identify where the attack could have been interrupted. For example, reporting the suspicious email before clicking it or notifying IT immediately after entering credentials could have allowed the incident to be contained before ransomware was deployed.

V. DISCUSSION

This section discusses the current status of the developed model, its limitations, and the planned next steps. The aim is to critically assess how well the proposed approach addresses the practical needs of healthcare environments and to identify remaining gaps before broader implementation. In addition, the section outlines concrete actions to further develop, evaluate, and integrate the model into the broader XiA project framework. The proposed model can be viewed as a prototype knowledge structure for future XiA persona-based assistants. The attack stages, explanations, and role-specific recommendations could be integrated into an intelligent agent that provides context-aware cybersecurity guidance to healthcare professionals in real time. The most important limitation is that the model has not been tested with real clinical staff, so it is not yet known whether it changes actual behavior. Most of the research reviewed was conducted in Western Europe and North America, so the model may not apply equally well to other healthcare systems. The XiA persona was built from published research rather than interviews with real nurses, and the training scenario is simpler than a real targeted phishing attack, which is usually harder to detect. The attack-chain model, XiA persona, role-based action table, training

scenario, and literature review have been completed. Future work includes developing additional personas, producing supporting learning materials, and carrying out the planned evaluation. To support the further development of the model and its implementation, the following concrete actions are proposed:

- Expand the persona set to cover the three XiA learner groups (Developers, Implementers, and End-Users) by developing representative personas, role-specific phishing scenarios, and tailored response guidance for each group.
- Produce a printable checklist for hospital workstations.
- Run a tabletop exercise with peers. A tabletop exercise is a structured group discussion in which participants talk through a simulated attack scenario to assess whether the model is clear and easy to follow.
- Write up the final findings and include them in the XiA project documentation.

VI. CONCLUSION

Cybersecurity training is often designed from an organizational or technical perspective and may provide limited support for frontline healthcare professionals when making security-related decisions in everyday clinical practice. This Work-in-Progress study presents the Role-Based Phishing-to-Ransomware Readiness Model, which combines a healthcare-oriented phishing-to-ransomware attack chain, a XiA-inspired persona-based learning approach, and role-based response guidance for clinical staff. Rather than proposing a new cybersecurity framework or attack taxonomy, the model adapts established cybersecurity concepts into a form that is understandable and actionable for non-technical healthcare professionals. The work contributes a preliminary design artifact that may help bridge the gap between organizational cybersecurity guidance and the practical realities of clinical work. The proposed model has been developed through literature review, Design Science Research principles, and feedback from XiA-related experts. However, the model has not yet been evaluated with healthcare professionals in operational settings. Consequently, no conclusions can currently be drawn regarding its effectiveness in improving phishing awareness, reporting behavior, decision-making, or ransomware readiness. Despite these limitations, the study suggests that persona-based and role-oriented approaches may offer a useful direction for translating cybersecurity knowledge into practical guidance for different healthcare stakeholder groups. The model may also provide a foundation for future XiA-compatible intelligent assistants that deliver explainable, context-aware cybersecurity support. Future work will focus on empirical validation with healthcare professionals, refinement of the personas and learning materials, and the development of measurable evaluation criteria. Such studies are needed to determine whether the proposed approach can contribute to improved cybersecurity behavior and preparedness in real healthcare environments.

ACKNOWLEDGEMENTS

The authors would like to thank the XiA project (funded by the European Union under Grant Agreement No. 101187650) for providing an inspiring and timely research topic, as well as for the valuable guidance and support throughout the research process. The constructive feedback helped shape the model's development and improve its practical relevance.

ETHICS DECLARATION

This project did not involve experiments with human participants or the collection or processing of personal data. No ethical clearance was required. All research followed the academic integrity guidelines of Laurea University of Applied Sciences.

AI DECLARATION

Artificial intelligence tools, including ChatGPT and Claude, were used to support language editing, structure checking, and clarity improvements during the preparation of this paper. The authors made all ideas, analyses, conclusions, and final editorial decisions.

REFERENCES

[1]. 'Cyber attacks on London's hospitals affect 800 planned operations'. Accessed: Jun. 05, 2026. [Online]. Available: https://www.bbc.com/news/articles/cd11v377eywo

[2]. P. K. Yeng, M. A. Fauzi, B. Yang, and P. Nimbe, 'Investigation into Phishing Risk Behavior among Healthcare Staff', Information, vol. 13, no. 8, p. 392, Aug. 2022, doi: 10.3390/info13080392.

[3]. NIST, 'Cybersecurity Framework', NIST, Nov. 2013, Accessed: Dec. 01, 2025. [Online]. Available: https://www.nist.gov/cyberframework

[4]. ENISA, 'European Cybersecurity Skills Framework (ECSF) | ENISA'. Accessed: Dec. 01, 2025. [Online]. Available: https://www.enisa.europa.eu/topics/skills-and-competences/skills-development/european-cybersecurity-skills-framework-ecsf

[5]. M. I. Mashinchi, T. Acton, and P. M. Datta, 'When healthcare becomes sick: Recovering from ransomware', J. Inf. Technol. Teach. Cases, vol. 16, no. 1, pp. 92–101, May 2026, doi: 10.1177/20438869241279443.

[6]. P. Rathod et al., 'Cybersecurity Practical Skills Gaps in Europe: Market Demand and Analysis', 2023.

[7]. J. Prümmer, T. van Steen, and B. van den Berg, 'Assessing the effect of cybersecurity training on End-users: A Meta-analysis', Comput Secur, vol. 150, no. C, maaliskuu 2025, doi: 10.1016/j.cose.2024.104206.

[8]. D. Hillman, Y. Harel, and E. Toch, 'Evaluating organizational phishing awareness training on an enterprise scale', Comput. Secur., vol. 132, p. 103364, Sep. 2023, doi: 10.1016/j.cose.2023.103364.

[9]. R. van Bavel, N. Rodríguez-Priego, J. Vila, and P. Briggs, 'Using protection motivation theory in the design of nudges to improve online security behavior', Int. J. Hum.-Comput. Stud., vol. 123, pp. 29–39, Mar. 2019, doi: 10.1016/j.ijhcs.2018.11.003.

[10]. M. Syafrizal, S. R. Selamat, and N. Zakaria, 'Analysis of Cybersecurity Standard and Framework Components', Int. J. Commun. Netw. Inf. Secur., vol. 12, pp. 417–432, Dec. 2020, doi: 10.17762/ijcnis.v12i3.4817.

[11]. C. S. Kruse, B. Frederick, T. Jacobson, and D. K. Monticone, 'Cybersecurity in healthcare: A systematic review of modern threats and trends', Technol. Health Care, vol. 25, no. 1, pp. 1–10, Jan. 2017, doi: 10.3233/THC-161263.

[12]. XiA, 'XiA – Xpanding Innovative Alliance'. Accessed: Nov. 12, 2025. [Online]. Available: https://xia-project.iscte-iul.pt/

[13]. 'Shaping the Future of Digital Health: Understanding XiA Learner Groups – XiA'. Accessed: Jun. 30, 2026. [Online]. Available: https://xia-project.iscte-iul.pt/shaping-the-future-of-digital-health-understanding-xia-learner-groups/

[14]. A. R. Hevner, S. T. March, J. Park, and S. Ram, 'Design Science in Information Systems Research1', Manag. Inf. Syst. Q., vol. 28, no. 1, pp. 75–106, Mar. 2004, doi: 10.2307/25148625.

[15]. Eric Hutchins, Michael Cloppert, and Rohan Admin, 'Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains', in Leading Issues in Information Warfare and Security Research, vol. One, Academic Conferences Limited, 2011, pp. 80–106.

[16]. B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, and C. B. Thomas, 'MITRE ATT&CK®: Design and Philosophy'.

[17]. 'Cyber Hygiene in the Health Sector | ENISA'. Accessed: Jun. 05, 2026. [Online]. Available: https://www.enisa.europa.eu/publications/cyber-hygiene-in-the-health-sector

[18]. 'Cybersecurity Framework', NIST, Nov. 2013, Accessed: Jun. 05, 2026. [Online]. Available: https://www.nist.gov/cyberframework

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Explainable Machine Learning for Heart Disease Risk Prediction: Comparative Evaluation of ML Algorithms, Feature Importance Stability, and Clinical Interpretability

Medina Ferhatović
Dept. of Software Engineering
University of Zenica
Zenica, Bosnia and Herzegovina
ferhatovic.medina.21@size.ba

Belmin Haračić
Dept. of Software Engineering
University of Zenica
Zenica, Bosnia and Herzegovina
haracic.belmin.21@size.ba

Narcisa Hadžajlić
Dept. of Software Engineering
University of Zenica
Zenica, Bosnia and Herzegovina
narcisa.hadzajlic@dl.unze.ba

Abstract
This paper presents an explainable machine learning framework for predicting heart disease risk that evaluates models not only on predictive performance but also on the stability and clinical relevance of their explanations. Two datasets with different characteristics were used: the CDC dataset (over 440,000 records and 39 features after encoding) and the smaller clinical Heart Failure dataset (918 patients and 11 features). Six machine learning algorithms were implemented and compared: Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and TabNet. Models were evaluated using standard classification metrics, with an emphasis on recall given the clinical cost of false negatives. SHAP, LIME, Permutation Feature Importance, and Partial Dependence Plots were employed to explain model predictions. The stability of SHAP explanations was assessed across five different random seeds and five independent train/test splits using the Jaccard and Spearman coefficients. At the same time, clinical interpretability was evaluated by comparing the most important features with established risk factors reported in clinical guidelines. A composite score was proposed to integrate predictive performance, explanation stability, and clinical alignment into a single evaluation metric. The results demonstrate that the model achieving the highest ROC-AUC does not necessarily provide the most stable or clinically aligned explanations, indicating that model selection for medical applications should not rely solely on conventional predictive performance metrics.

Keywords: explainable artificial intelligence, heart disease prediction, SHAP, feature importance stability, clinical interpretability, machine learning

I. INTRODUCTION

Cardiovascular diseases are the leading cause of mortality worldwide, accounting for nearly one-third of all annual deaths, many of which are preventable if high-risk patients are identified at an early stage. Machine learning (ML) algorithms have demonstrated increasingly reliable predictive performance in disease detection, with ensemble models such as Random Forest, XGBoost, and gradient-boosted variants consistently achieving the highest classification accuracy when trained on clinical and demographic data [1],[2],[3]. However, these high-performing models are inherently black-box systems whose decision-making processes cannot be directly interpreted [4],[5]. In medical practice, this represents a significant limitation, as clinicians who cannot understand why a patient has been classified as high risk may find it difficult to justify such predictions within clinical decision-making protocols [6],[7]. To bridge the gap between predictive performance and trustworthiness, Explainable Artificial Intelligence (XAI) techniques, particularly SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), are widely employed to quantify the contribution of individual features to model predictions [4],[6],[8],[9],[10].



This study is based on the premise that, in medical applications, the best model is not necessarily the one with the highest predictive accuracy, but rather the one that achieves the best balance between predictive performance, explanation stability, and clinical relevance. To address this issue, an explainable machine learning framework is developed and evaluated, assessing models across all three dimensions simultaneously. The authors are not medical professionals; clinical interpretability here refers solely to alignment with risk factors from published guidelines [1],[6],[7],[8], not independent medical validation.

II. LITERATURE REVIEW

A considerable body of research has focused on ML algorithms for heart disease prediction, with the most commonly compared models including Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Gradient Boosting, and XGBoost [1],[2],[3]. Ensemble methods consistently outperform traditional linear models in predictive performance [2]. To overcome the black-box nature of complex predictive models, many recent studies combine heart disease prediction with Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME [4–11]. Fewer studies investigate the stability of feature importance analyses and show that feature rankings may vary across different data samples or model initializations [12],[13],[14]. Based on the reviewed literature, three main research directions can be identified. The first improves predictive performance [1], [2], [3]; the second applies XAI for transparency [4–11]; and the third investigates explanation stability [12], [13], [14], but independently of heart disease prediction. None of the reviewed studies simultaneously considers all three aspects: predictive performance, the stability of XAI explanations across different data splits, random seed values, and the validation of the clinical relevance of these explanations against established clinical guidelines. Addressing this research gap is the primary objective of this study, which integrates these three dimensions into a unified framework for evaluating machine learning models.

III. METHODOLOGY

Previous studies on machine learning and XAI for heart disease prediction recommend comparing multiple machine learning models using standardized preprocessing procedures and publicly available datasets while complementing predictive evaluation with XAI techniques such as SHAP and LIME to improve model transparency [1–11]. Moreover, recent research highlights the need to assess the reliability and consistency of feature importance explanations [12–14]. Based on these findings, this study applies a comparative experimental methodology using two benchmark datasets, six machine learning models, multiple XAI techniques, and an explainability stability analysis.

Datasets: Two publicly available datasets from the Kaggle platform were used in this study [15],[16]. These datasets differ in size, data type, and clinical specificity.

Dataset A, Indicators of Heart Disease [15], contains over 440,000 CDC BRFSS survey records with demographic, lifestyle, and medical condition variables. The target variable (HadHeartAttack) is highly imbalanced (94.4% negative, 5.6% positive)

Dataset B, Heart Failure Prediction Dataset [16], combines five clinical datasets into 918 patient records with 11 clinical variables (age, sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope). Classes are approximately balanced.

Data Preprocessing: Numerical variables with missing values were imputed using the median, while categorical variables were imputed using the mode. Categorical features were encoded, and numerical features were

standardized using the StandardScaler transformation. The dataset was split into training and testing sets in an 80/20 ratio with stratification of the target variable. To prevent data leakage, scaling and SMOTE (Synthetic Minority Over-sampling Technique) [17] were applied exclusively on the training set.

Implemented Machine Learning Models: Six algorithms were selected based on prevalence in the literature and methodological complementarity: Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, and TabNet [1],[2],[3],[4].

Evaluation Metrics: Models were evaluated using accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, with emphasis on recall and PR-AUC given the clinical cost of false negatives.

XAI Methods: To interpret model predictions, four complementary Explainable Artificial Intelligence (XAI) techniques were applied: SHAP, LIME, Permutation Feature Importance (PFI), and Partial Dependence Plots (PDP).

Explainability Stability Analysis: A central contribution of this study is the evaluation of SHAP explanation stability at two levels: five random seeds (seeds 42–46, fixed data split) and five independent 80/20 train-test splits, using Jaccard@10 and Spearman ρ as stability metrics.

Clinical Interpretability: Clinical interpretability was evaluated by comparing the top-10 SHAP features with established cardiovascular risk factors reported in clinical guidelines.

Composite Scoring: A composite score integrating predictive performance, explanation stability, and clinical alignment was computed to compare models.

IV. RESULTS

Table 1 presents the performance of all models on Dataset A. Due to class imbalance, accuracy is misleading on Dataset A; values above 0.94 reflect dominance of the majority class. The three gradient boosting models (XGBoost, LightGBM, and CatBoost) achieve the highest ROC-AUC (≈ 0.886) and PR-AUC (0.41–0.42), but exhibit very low recall for the positive class (0.237–0.247), meaning that approximately three-quarters of actual positive cases are missed. In contrast, Logistic Regression and TabNet, despite slightly lower ROC-AUC values (0.878 and 0.862, respectively), achieve substantially higher recall (0.765 and 0.736) and balanced accuracy above 0.78. This inversion is a direct consequence of decision-threshold effects and SMOTE balancing, and is clinically significant because it reflects the proportion of correctly identified patients with heart disease.

Table 1. Model Performance - Dataset A

Model	**Model Performance - Dataset A**					
	Recall (HD)	***F1 (macro)***	***Balanced Accuracy***	***ROC-AUC***	***PR-AUC***	***Specificity***
Logistic Regression	0.765	0.612	0.795	0.878	0.391	0.826
Random Forest	0.438	0.710	0.704	0.879	0.380	0.970
XGBoost	0.247	0.659	0.618	0.887	0.411	0.989
LightGBM	0.237	0.653	0.613	0.886	0.411	0.989
CatBoost	0.241	0.658	0.616	0.886	0.420	0.990
TabNet	0.736	0.605	0.780	0.862	0.347	0.823

Model	Model Performance - Dataset A					
	Recall (HD)	***F1 (macro)***	***Balanced Accuracy***	***ROC-AUC***	***PR-AUC***	***Specificity***
Logistic Regression	0.765	0.612	0.795	0.878	0.391	0.826
Random Forest	0.438	0.710	0.704	0.879	0.380	0.970
XGBoost (tuned)	0.472	0.718	0.720	0.881	0.395	0.968
CatBoost (tuned)	0.471	0.781	0.719	0.880	0.405	0.967

On the smaller, balanced Dataset B, all models achieve consistently high performance (Table 2), with ROC-AUC ranging from 0.897 to 0.933 and recall from 0.853 to 0.941. Unlike Dataset A, there is no accuracy–recall trade-off, confirming that class imbalance, and not algorithm choice, drives model behavior on imbalanced data.

Table 2. Model Performance — Dataset B

Model	Model Performance - Dataset B					
	Recall (HD)	***F1 (macro)***	***Balanced Accuracy***	***ROC-AUC***	***PR-AUC***	***Specificity***
Logistic Regression	0.931	0.866	0.862	0.897	0.879	0.793
Random Forest	0.922	0.901	0.900	0.927	0.914	0.878
XGBoost	0.853	0.858	0.859	0.926	0.934	0.866
LightGBM	0.873	0.874	0.875	0.932	0.936	0.878
CatBoost	0.863	0.874	0.876	0.931	0.935	0.890
TabNet	0.931	0.866	0.862	0.933	0.944	0.793
XGBoost (tuned)	0.922	0.878	0.875	0.921	0.917	0.829
CatBoost (tuned)	0.941	0.894	0.891	0.931	0.931	0.842

1. Hyperparameter Tuning

Bayesian optimization using the Optuna library [18] was applied to XGBoost and CatBoost, the two models achieving the highest macro F1-score on Dataset A. The optimization process significantly improved the F1 (macro) score from a baseline of approximately 0.658 to approximately 0.718 for both models, primarily by increasing recall for the positive class (HD), from approximately 0.24 to approximately 0.47.

2. Global Feature Importance (SHAP)

Global SHAP analysis reveals a high degree of consistency across models. On Dataset A, HadAngina and AgeCategory are the most dominant features across all models, followed by GeneralHealth, SleepHours, and RemovedTeeth (Figure 1). On Dataset B, ST_Slope emerges as the most influential feature across all models (Mean |SHAP| range: 1.15–2.05), followed by ChestPainType and Cholesterol (Figure 2). Pairwise Spearman rank correlations between the five models' full SHAP-based feature rankings quantify this consistency: the four tree-based models (Random Forest, XGBoost, LightGBM, and CatBoost) show strong mutual agreement (ρ = 0.88–0.97 on Dataset A, ρ = 0.78–0.95 on Dataset B), whereas Logistic Regression correlates more moderately with the tree ensembles (ρ = 0.65–0.73 on Dataset A, ρ = 0.69–0.78 on Dataset B), reflecting the expected divergence between linear and non-linear explanation structures. The dominance of ST-segment slope is clinically consistent with its well-established prognostic importance in ECG-based cardiac assessment.

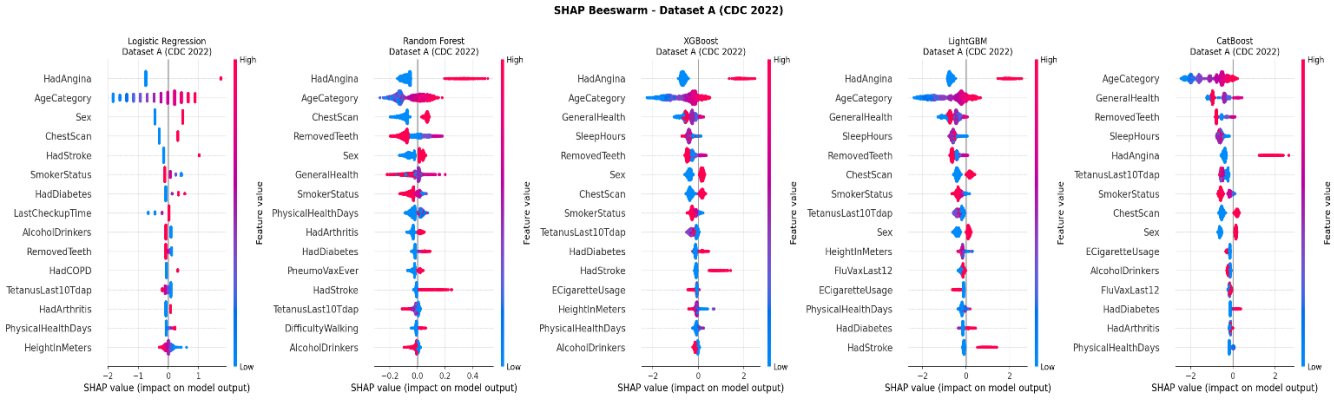


Figure 1. SHAP beeswarm plot for Dataset A. Red points (high feature values) for *HadAngina* consistently shift the prediction toward the positive class.

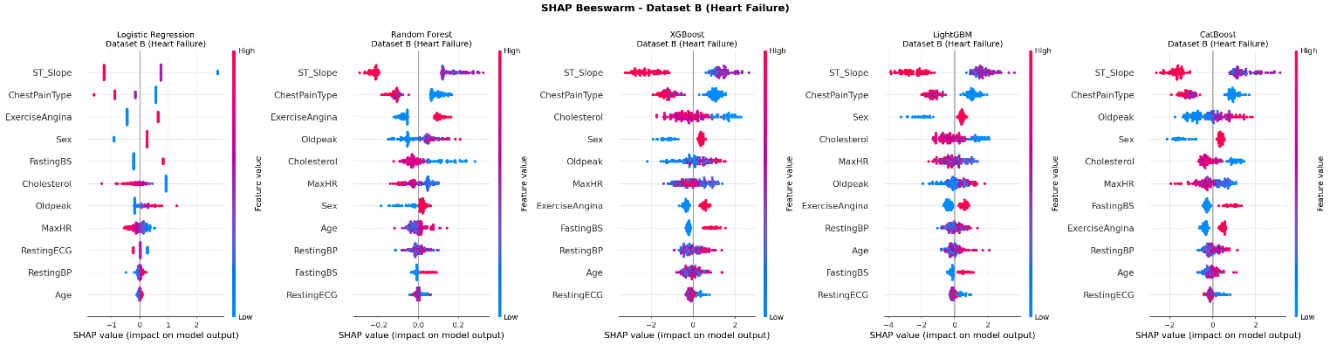


Figure 2. SHAP beeswarm plot for Dataset B. Red points (high feature values) for *ST_Slope* consistently shift the prediction toward the positive class.

3. Local Explanations (SHAP, LIME)

Local SHAP and LIME explanations (Figure 3) were generated for four representative cases (TP, FN, TN, FP). For correctly classified positives in Dataset B, ST_Slope, Cholesterol, and FastingBS consistently support the positive prediction, whereas for false negatives, they have the opposite effect, indicating atypical patient profiles. The agreement between SHAP and LIME further validates the identified factors.

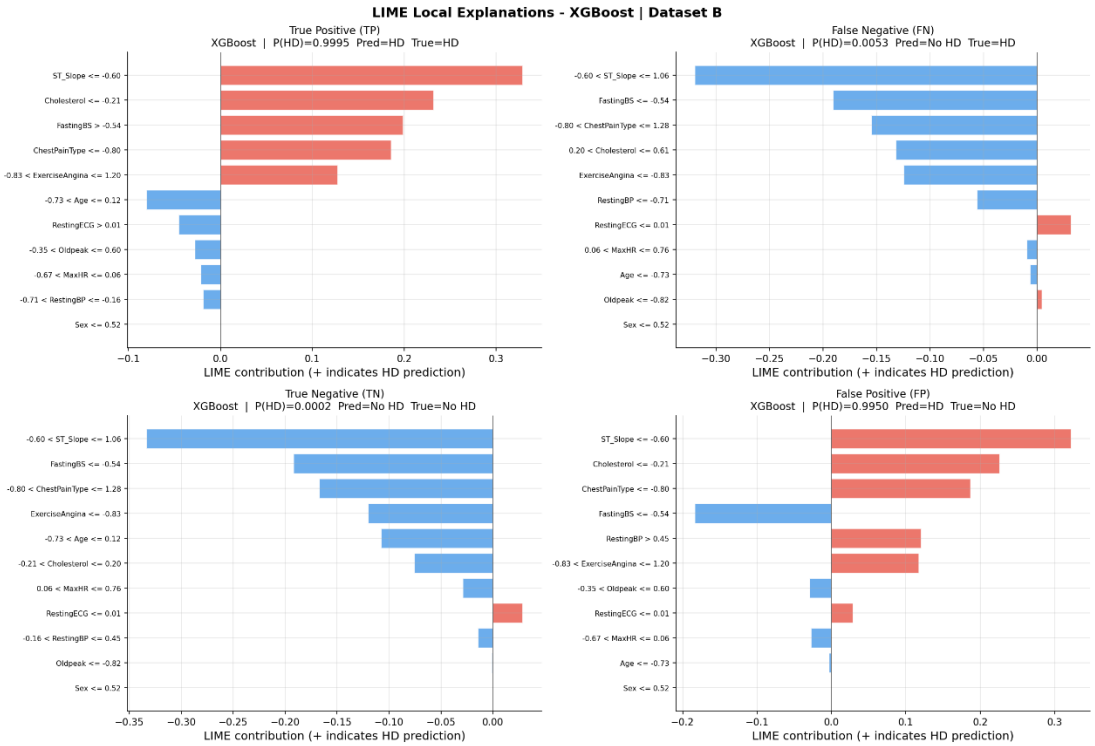


Figure 3. LIME local explanations (XGBoost, Dataset B) for TP, FN, TN, and FP cases.

4. Permutation Feature Importance

Permutation Feature Importance (PFI) confirms SHAP rankings: HadAngina and AgeCategory dominate on Dataset A, while ST_Slope clearly leads on Dataset B. Convergence of both methods strengthens confidence in the identified factors. Quantitatively, the top-15 feature sets identified by SHAP and PFI for the best-performing model (XGBoost) show a Jaccard similarity of 0.667 and a Spearman rank correlation of ρ = 0.851 on Dataset A, and a perfect Jaccard similarity of 1.000 with ρ = 0.818 on Dataset B, providing independent, model-agnostic corroboration of the SHAP-based feature rankings.

5. Explanation Stability

Table 3 summarizes the stability analysis of SHAP explanations for XGBoost. Across five random seeds and five independent train-test splits, the model achieves a Jaccard@10 score of 0.891 and a Spearman correlation coefficient of ρ = 0.988 on Dataset A, and a Jaccard@10 score of 1.000 and a Spearman ρ = 0.921 on Dataset B. To assess whether this stability generalizes beyond a single algorithm, the same protocol was subsequently applied to all five SHAP-compatible models (XGBoost, LightGBM, CatBoost, Random Forest, and Logistic Regression); TabNet was excluded, as it lacks SHAP TreeExplainer support. Across both the seed-based and split-based analyses, all five models exhibit comparably high stability, with Jaccard@10 ranging from 0.855 to 1.000 and Spearman ρ from 0.973 to 0.998 on Dataset A, and Jaccard@10 from 0.855 to 1.000 and Spearman ρ from 0.912 to 1.000 on Dataset B. This confirms that the stability of SHAP explanations is not an artifact specific to XGBoost, but a consistent property shared by both tree-based ensembles and the linear baseline.

Table 3. Shap Explanation Stability (Xgboost)

Level of Analysis	**SHAP Explanation Stability (XGBoost)**		
	Dataset	***Jaccard@10***	***Spearman ρ***
5 random seeds	A	0.891	0.988
5 train/test split	A	0.891	0.988
5 random seeds	B	1.000	0.921
5 train/test split	B	1.000	0.921

Extended results for all five models are summarized in the text above; XGBoost is retained in Table 3 as the representative case.

6. Clinical Interpretability

A comparison of the top-10 SHAP features with risk factors in medical guidelines shows strong alignment (Figure 4). On Dataset A, the proportion of clinically expected features ranges from 70% (LightGBM, CatBoost) to 90% (Random Forest). On Dataset B, all models achieve 100%, as expected, since it is constructed entirely from clinically relevant variables.

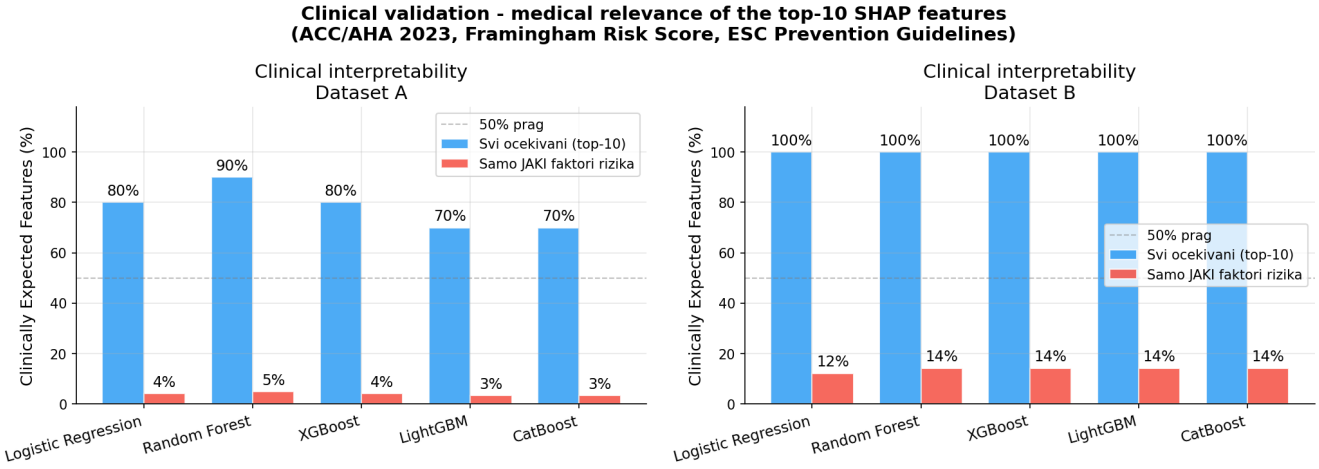


Figure 4. Clinical alignment of top-10 SHAP features with medical guidelines for both datasets.

7. Final Synthesis – Composite Score

Table 4 presents the computed composite scores. On Dataset A, Random Forest achieves the highest composite score (C = 0.852) despite lower ROC-AUC than gradient boosting models, due to superior clinical alignment (90%) and balanced recall. On Dataset B, all models perform similarly (C = 0.932–0.951), with Random Forest leading.

Table 4. Composite Score

Model	Composite Score				
	ROC-AUC	***F1m***	***Jaccard***	***Klin.%***	***C***
DATASET A					
Logistic Regression	0.878	0.612	0.891	0.80	0.812
Random Forest	0.879	0.710	0.891	0.90	**0.852**
XGBoost	0.887	0.659	0.891	0.80	0.825
LightGBM	0.886	0.653	-	0.70	0.803
CatBoost	0.886	0.658	-	0.70	0.804
TabNet	0.862	0.605	-	-	0.744
DATASET B					
Logistic Regression	0.897	0.866	0.691	1.00	0.932
Random Forest	0.927	0.901	1.000	1.00	**0.951**
XGBoost	0.926	0.858	1.000	1.00	0.942
LightGBM	0.932	0.874	-	1.00	0.948
CatBoost	0.931	0.874	-	1.00	0.947
TabNet	0.933	0.866	-	-	0.847

V. DISCUSSION

The central finding of this study is that the model achieving the highest ROC-AUC is not necessarily the one that provides the best overall balance for clinical application. On Dataset A, gradient boosting models (XGBoost, LightGBM, CatBoost) achieve a marginally higher ROC-AUC (≈0.886) than Random Forest (0.879), but with a recall for the positive class of only ≈0.24, meaning that in practice they would miss approximately three-quarters of truly affected patients.

The contrast between the two datasets further highlights the role of class imbalance. On the balanced Dataset B, all models achieve simultaneously high accuracy and high recall, and differences in composite scores are minimal (0.932–0.951). The sharp inverse relationship between accuracy and recall observed in Dataset A is therefore not an inherent property of the algorithms, but a consequence of severe class imbalance. SHAP explanations proved both stable (Jaccard@10 0.891–1.000, Spearman $\rho > 0.92$) and clinically meaningful: HadAngina, AgeCategory, and GeneralHealth (Dataset A) and ST_Slope, ChestPainType, and Cholesterol (Dataset B) all align with established cardiovascular risk factors [1],[6],[7],[8], further confirmed by PFI convergence.

The study has several limitations. First, the authors are not medical professionals; therefore, clinical alignment refers strictly to agreement with risk factors reported in published literature and guidelines, rather than independent medical validation. Second, both datasets originate from specific geographic and population contexts (CDC BRFSS – USA; combined clinical datasets – Europe/USA), so generalization to other populations requires further validation. Finally, the weighting scheme used in the composite score (0.40/0.20/0.20/0.20) reflects a prioritization of predictive performance while treating the other dimensions equally. However, it can be adjusted to suit specific clinical requirements.

VI. CONCLUSION AND FUTURE WORK

This study presented an explainable ML framework for heart disease prediction and evaluated it in terms of predictive performance, explanation stability, and clinical alignment. On the imbalanced Dataset A, Random Forest achieved the best composite score (C=0.852) despite lower ROC-AUC than gradient boosting models, which missed ~75% of positive cases. SHAP explanations were highly stable (Jaccard@10 0.891–1.000) and aligned with established medical risk factors (70–100%).

The results indicate that the model achieving the highest ROC-AUC does not necessarily provide the most stable or clinically aligned explanations. Future work will explore alternative balancing techniques (ADASYN, class weights), validate across diverse populations, and develop a clinical decision-support prototype in collaboration with medical experts.

REFERENCES

[1]. R. Kumar, S. Garg, R. Kaur, M. G. M. Johar, S. Singh, S. V. Menon, P. Kumar, A. M. Hadi, S. A. Hasson, and J. Lozanović, "A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions," Front. Artif. Intell., vol. 8, article no. 1583459, May 2025, doi: 10.3389/frai. 2025.1583459.

[2]. M. D. Teja and G. M. Rayalu, "Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance," BMC Cardiovasc. Disord., vol. 25, article no. 212, Mar. 2025, doi: 10.1186/s12872-025-04627-6.

[3]. G. Gul, I. A. Korejo, D. N. Hakro, H. Alqahtani, A. Abbasi, M. Babar, O. Al Rahbi, and N. I. Ali, "Machine learning and ensemble methods for cardiovascular disease prediction: a systematic review of approaches, performance trends, and research challenges," Computers, vol. 15, no. 1, article no. 25, Jan. 2026, doi: 10.3390/computers15010025.

[4]. S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," Sci. Rep., vol. 15, article no. 13912, Apr. 2025, doi: 10.1038/s41598-025-97547-6.

[5]. A. S. Ningthoujam, S. Sharma, and A. Nandi, "Explainable AI-based coronary heart disease prediction: enhancing model transparency in clinical decision making," bioRxiv, Mar. 2025, doi: 10.1101/2025.03.20.644456.

[6]. A. Bilal, A. Alzahrani, K. Almohammadi, M. Saleem, M. S. Farooq, and R. Sarwar, "Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease," Front. Med., vol. 12, article no. 1596335, Jul. 2025, doi: 10.3389/fmed.2025.1596335.

[7]. N. Kailasanathan, G. Ezhilarasan, S. Selvarajan, R. K. Dhanaraj, D. Pamucar, and N. Shankar, "Heart disease prediction with a feature-sensitized interpretable framework for the Internet of Medical Things sensors," Front. Digit. Health, vol. 7, article no. 1612915, Oct. 2025, doi: 10.3389/fdgth.2025.1612915.

[8]. Md. A. Talukder, A. S. Talaat, M. Kazi, and A. Khraisat, "XAI-HD: an explainable artificial intelligence framework for heart disease detection," Artif. Intell. Rev., vol. 58, article no. 385, Dec. 2025, doi: 10.1007/s10462-025-11385-6.

[9]. M. Waqar, M. B. Shahnawaz, S. Saleem, H. Dawood, U. Muhammad, and H. Dawood, "Enhancing heart attack prediction: feature identification from multiparametric cardiac data using explainable AI," Algorithms, vol. 18, no. 6, article no. 333, Jun. 2025, doi: 10.3390/a18060333.

[10]. N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, "XAI-augmented voting ensemble models for heart disease prediction: a SHAP and LIME-based approach," Bioengineering, vol. 11, no. 10, article no. 1016, Oct. 2024, doi: 10.3390/bioengineering11101016.

[11]. K. K. Napa, R. Govindarajan, S. Sathya, J. S. Murugan, and B. K. P. Vijayammal, "Comparative analysis of explainable machine learning models for cardiovascular risk stratification using clinical data and Shapley additive explanations," Smart Health, 2025, doi: 10.1016/j.smhl.2025.100900.

[12]. J. Goldwasser and G. Hooker, "Statistical significance of feature importance rankings," in Proc. Conf. Uncertainty Artif. Intell. (UAI), 2024, doi: 10.48550/arXiv.2401.15800.

[13]. A. Xue, R. Alur, and E. Wong, "Stability guarantees for feature attributions with multiplicative smoothing," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023, doi: 10.48550/arXiv.2307.05902

[14]. N. Harel, R. Gilad-Bachrach, and U. Obolski, "Inherent inconsistencies of feature importance," arXiv, 2022, doi: 10.48550/arXiv. 2206.08204.

[15]. K. Pytlak, "Indicators of heart disease (2022 update)," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[16]. Fedesoriano, "Heart failure prediction dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[17]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair. 953.

[18]. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2019, pp. 2623–2631, doi: 10.1145/3292500.3330701.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Preliminary Development of an Aircraft Structural Damage Classification System Using Deep Learning

Reynaldo Ángel Serrano Cruz
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
Tecnológico Nacional de México (TECNM)
Cuernavaca, Mor. México
d22ce077@cenidet.tecnm.mx
ORCID:0009-0002-6563-8358

Gloria Lilia Osorio Gordillo
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
Tecnológico Nacional de México (TECNM)
Cuernavaca, Mor. México
gloria.og@cenidet.tecnm.mx
ORCID:0000-0002-3445-8867

Rodolfo Amalia Vargas Méndez
Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)
Tecnológico Nacional de México (TECNM)
Cuernavaca, Mor. México
rodolfo.vm@cenidet.tecnm.mx
ORCID: 0000-0003-4915-9787

Abstract
Aircraft structural damage inspection remains a critical task for ensuring operational safety and airworthiness. Conventional inspection procedures still rely heavily on human visual assessment, making them susceptible to fatigue, subjectivity, and inconsistent damage characterization. Recent advances in deep learning have demonstrated significant potential for automating visual inspection processes. This work presents the preliminary development of an aircraft structural damage classification system based on convolutional neural networks and transfer learning. A ResNet50 architecture was implemented as the base classifier in a cloud-based environment and trained to identify four structural condition categories: crack, dent, corrosion, and no-damage. Image preprocessing and data augmentation techniques were applied to improve training robustness and generalization capability. Preliminary results indicate that the proposed approach can successfully learn visual patterns associated with the evaluated structural conditions. The proposed classifier achieved a test accuracy of 96.20% and a weighted F1-score of 0.9617 on an independent test subset of 79 images. However, isolated misclassifications remained between visually similar defects such as corrosion and cracks. These results support the feasibility of the proposed methodology while highlighting the need for dataset expansion and further model optimization. The developed framework provides a foundation for future work involving damage localization, geometric measurement, explainable artificial intelligence, and integration with aircraft maintenance documentation.

Keywords: Aircraft structural inspection, Deep Learning, Convolutional Neural Networks, Transfer Learning, ResNet50, Damage Classification, Computer Vision, Aircraft Maintenance

I. INTRODUCTION

Aircraft structural inspection plays a fundamental role in maintaining operational safety and ensuring the airworthiness of commercial aviation fleets [1], [2]. Aircraft structures are continuously exposed to operational loads, environmental conditions, impacts, corrosion, and material fatigue, which can lead to structural defects such as cracks, dents, perforations, and surface degradation. If such damage is not detected and properly characterized during maintenance procedures, structural integrity and flight safety may be compromised. Despite technological advances in aviation maintenance, visual inspection performed by human operators remains one of the most widely used methods for structural evaluation. However, manual inspection procedures are inherently dependent on inspector experience [1], [2], environmental conditions, fatigue, and subjective interpretation. These limitations may lead to inconsistencies in damage identification and classification, especially when small defects or complex surface patterns are involved. Recent developments in artificial intelligence and computer vision have enabled the application of deep learning techniques to automated defect

detection and classification tasks [3], [4]. Convolutional Neural Networks (CNNs) have demonstrated strong performance in image recognition applications by automatically extracting hierarchical visual features from complex datasets. Consequently, several researchers have explored the use of deep learning architectures such as ResNet, YOLO, Faster R-CNN, and Mask R-CNN for structural inspection tasks in civil infrastructure and aerospace applications [5], [6]. In the aeronautical field, automated inspection systems have gained increasing attention due to their potential to reduce inspection times, improve repeatability, and assist maintenance personnel during decision-making processes [1], [5], [6], [7]. Previous studies have proposed approaches for crack detection, corrosion identification, aircraft skin inspection, and defect localization using both ground-based and drone-assisted image acquisition systems [4], [7], [8]. Nevertheless, many existing approaches remain focused on isolated tasks such as classification or detection. At the same time, limited work has explored integrated intelligent inspection frameworks that evolve toward technical interpretation and explainability [5]. This research presents the preliminary development of a structural damage classification system for aircraft surfaces using deep learning and transfer learning techniques. The current work corresponds to the initial classification stage of a broader modular architecture intended to support future functionalities, including damage localization, geometric estimation, explainable analysis, and rule-based interpretation derived from aircraft maintenance documentation. The main objective of this preliminary stage is to validate the feasibility of implementing a convolutional neural network to identify different categories of structural damage patterns in aircraft imagery. The proposed methodology employs a ResNet50 architecture trained via transfer learning and evaluated through initial classification experiments across multiple structural condition categories.

II. RELATED WORK

Aircraft structural inspection has traditionally relied on visual evaluation performed by trained maintenance personnel. Although widely adopted throughout the aviation industry, this approach may be affected by inspector experience, fatigue, environmental conditions, and subjective interpretation. To overcome these limitations, recent research has explored automated inspection systems based on computer vision and machine learning techniques that assist with damage identification and maintenance activities [9], [10]. The emergence of deep learning has significantly improved defect recognition performance. Convolutional Neural Networks (CNNs) have demonstrated strong capabilities in extracting hierarchical visual features from complex images and have been successfully applied to crack detection, corrosion assessment, and structural monitoring. Recent studies have also investigated data augmentation strategies to address the limited availability of aircraft defect imagery [11]. More recent studies have incorporated architectures such as U-Net, Mask R-CNN, Faster R-CNN, YOLO, and transformer-based models for damage detection and localization in both civil and aerospace structures [12], [6]. Despite these advances, most existing approaches focus primarily on detection or localization tasks. Fewer studies address multi-class aircraft damage classification as part of a broader inspection framework that can evolve toward geometric characterization, explainable analysis, and maintenance-oriented decision support. Motivated by this gap, the present work investigates the feasibility of a ResNet50-based classifier for distinguishing among crack, dent, corrosion, and no-damage conditions as the first stage of an intelligent aircraft structural inspection system.

III. METHODOLOGY

1. General System Architecture

The proposed methodology corresponds to the preliminary classification stage of a broader intelligent aircraft structural inspection framework. The current implementation focuses exclusively on the automatic classification of structural damage patterns from aircraft-related images.

The classification module receives an image of an aircraft surface and assigns it to one of four predefined categories: crack, dent, corrosion, or no-damage. The predicted class is intended to serve as the initial input for future modules involving damage localization, geometric characterization, and maintenance-oriented analysis. As illustrated in Figure 1, the processing pipeline consists of three main stages: image preprocessing, feature extraction using a pretrained ResNet50 backbone, and classification via a custom neural network head comprising Global Average Pooling, Batch Normalization, Dropout, and Softmax layers.

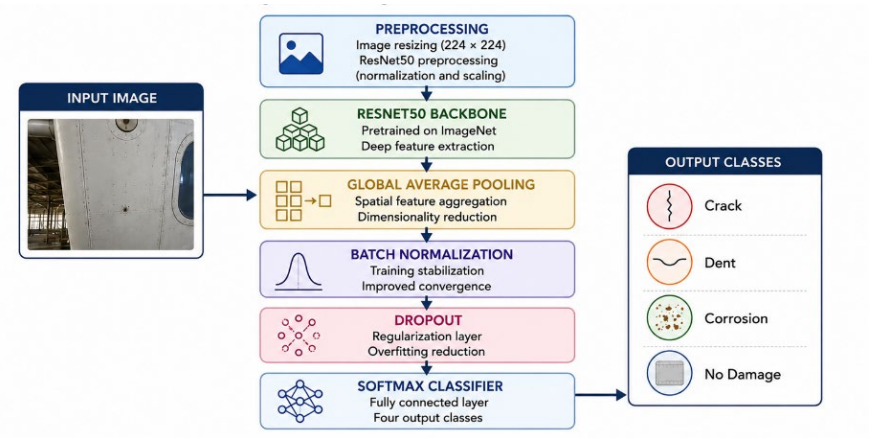


Figure 1. General architecture of the proposed aircraft structural damage classification system.

The classification function can be expressed as:

$$\hat{y} = f_\theta(x) \tag{1}$$

where x represents the input image, (f_θ)denotes the deep learning model with trainable parameters (θ) trainable parameters (θ), and ($\hat{y}$) corresponds to the predicted probability vector associated with the structural damage classes.

2. Image Preprocessing and Data Augmentation

All input images were resized to a spatial resolution of 224 × 224 pixels with three color channels, matching the input requirements of the pretrained ResNet50 architecture. Before training, images were processed using the standard ResNet50 preprocessing function, which applies the normalization strategy originally used during ImageNet training. To improve model robustness and reduce overfitting, data augmentation was applied exclusively to the training subset. The augmentation strategy included small rotations, horizontal and vertical translations, and zoom transformations, allowing the network to observe multiple variations of the same structural condition during training.

The augmentation process can be expressed as:

$$x' = T(x) \tag{2}$$

, ere (x)s the original image (T ·)epresents a random transformation operator, and (x′) is the augmented image used during training.

The implemented configuration employed a rotation range of 10°, width and height shifts of 0.10, and a zoom range of 0.10. Horizontal flipping was intentionally omitted because the orientation of structural damage may contain relevant visual information, depending on the inspected aircraft region.

3. Classification Model Based on ResNet50

The classification model was developed using the ResNet50 architecture as the backbone feature extractor. ResNet50 was selected due to its ability to learn deep hierarchical representations while mitigating the degradation problem commonly observed in very deep neural networks. The main principle of ResNet architectures is residual learning. Instead of directly learning a complete transformation, each residual block learns a residual mapping that is added to the input representation:

$$y(x) = F(x) + x \tag{3}$$

where x represents the input feature map, F(x) denotes the residual function learned by the convolutional layers, and y corresponds to the output of the residual block. In the proposed approach, the original fully connected layers of ResNet50 were removed, and the pretrained network was used as a deep feature extractor. The extracted feature maps were subsequently processed through a custom classification head comprising Global Average Pooling, Batch Normalization, Dropout, and a final Dense layer with a Softmax activation.

The Global Average Pooling operation reduces each feature map to a single representative value by averaging all spatial positions.

$$GAP_k = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} z_{i,j,k} \tag{4}$$

wher H and W are the eight and wdth o the feature map ,and $(z_{i,j,k})$orresponds to the activation value at position (i, j) of channel k. Batch Normalization was incorporated to stabilize training, while a Dropout layer with a rate of 0.30 was employed to reduce overfitting and improve generalization. The final classification stage was implemented using a Dense layer with Softmax activation, which transforms the network outputs into class probabilities:

$$P(y_i|x) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{5}$$

where $i. is\ the\ output\ logit\ associated\ with\ class\ i. and$ büyük $harf\ K\ represents$ the total number of classes. The predicted label corresponds to the class with the highest probability value.

4. Transfer Learning Strategy

Transfer learning was employed to leverage visual representations pre-trained on the ImageNet dataset. This approach is particularly advantageous when domain-specific datasets are limited, as demonstrated in previous

aerospace structural monitoring applications [4], [13]. The training process was performed in two stages. During the first stage, the ResNet50 backbone remained frozen and only the custom classification head was trained. This allowed the network to adapt its final decision layers to the aircraft damage classification task while preserving the generic visual features learned from ImageNet. In the second stage, a fine-tuning strategy was applied by selectively unfreezing the final layers of the ResNet50 backbone, excluding Batch Normalization layers. A reduced learning rate was used during this phase to refine high-level feature representations while minimizing the risk of degrading previously learned knowledge.

5. Training Configuration

The model was implemented in a cloud-based Python environment using TensorFlow and Keras. The input image size was set to 224 x to$(224 \text{ x } 224)$ pixels, and a batch size of 16 was used during training.

The initial training phase was configured for 20 epochs with an initial learning rate of:

$$\alpha = 1 \times 10^{-4} \tag{6}$$

For the fine-tuning phase, the model was trained for 10 additional epochs using a reduced learning rate of:

$$\alpha = 1 \times 10^{-5} \tag{7}$$

The Adam optimizer was used for parameter optimization, while class weights were incorporated to mitigate the existing class imbalance.

The model was trained with categorical cross-entropy as the loss function, which is well-suited to multiclass classification problems. The loss function is defined as:

$$\mathrm{L} = -\sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{y_i} \log(\hat{\mathrm{y}}_\mathrm{i}) \tag{8}$$

where (y_i)s the true one-ot encoded label for class (i), $(\hat{y}_i)$ is the probability predicted by the model, and (K) is the number of classes.

To address potential class imbalance, class weights were computed from the training set. The weight assigned to each class was obtained as:

$$\mathrm{w_i} = \frac{\mathrm{N}}{\mathrm{K} \cdot \mathrm{n_i}} \tag{9}$$

where (N) is the total number of training samples, (K) is the number of classes, and (n_i) is the number of samples belonging to class (i). These weights were incorporated during training to reduce bias toward classes with a larger number of samples. Classification accuracy was used as the primary metric for monitoring model performance during training.

IV. EXPERIMENTAL SETUP

The experimental dataset contained 794 aircraft-related images organized into four structural condition categories: dent, corrosion, crack, and no-damage. The dataset was assembled from publicly available aircraft imagery collected from multiple online and aeronautical sources and includes aircraft from the Airbus A320 family (A318, A319, A320, and A321). The collected images exhibit variations in viewing angle, illumination, background, and image resolution, reflecting the heterogeneous conditions of publicly available visual data. A stratified distribution was used to preserve the class proportions across training, validation, and test subsets. The resulting partition consisted of 555 training images, 160 validation images, and 79 independent test images, corresponding approximately to 69.9%, 20.2%, and 9.9% of the available dataset, respectively. Table I summarizes the distribution by class.

Table I. Dataset distribution across training, validation and test subsets.

Class	Train	Validation	Test	Total
Dent	14	5	2	21
Corrosion	170	49	24	243
Crack	191	55	27	273
No Damage	180	51	26	257
Total	555	160	79	794

Representative samples of the four structural conditions are shown in Figure 1. The experiments were conducted in Google Colab using TensorFlow and Keras with an NVIDIA Tesla T4 GPU (15 GB of GPU memory). A ResNet50 backbone pretrained on ImageNet was employed. The classification head was initially trained for 20 epochs with a learning rate of $1x10^{-4}$, followed by 10 fine-tuning epochs using an initial learning rate of $1x.10^{-5}$. A batch size of 16 was used throughout the experiments.



Figure 1. Representative examples of aircraft structural conditions included in the dataset.

V. RESULTS AND DISCUSSION

Figure 2 shows the evolution of training and validation accuracy and loss throughout the 30 training epochs. Both sets of curves exhibit a stable learning trend, with validation accuracy reaching approximately 95%. The relatively small gap between training and validation performance suggests that data augmentation, regularization, and fine-tuning helped limit overfitting, although a moderate gap persists in the final epochs.

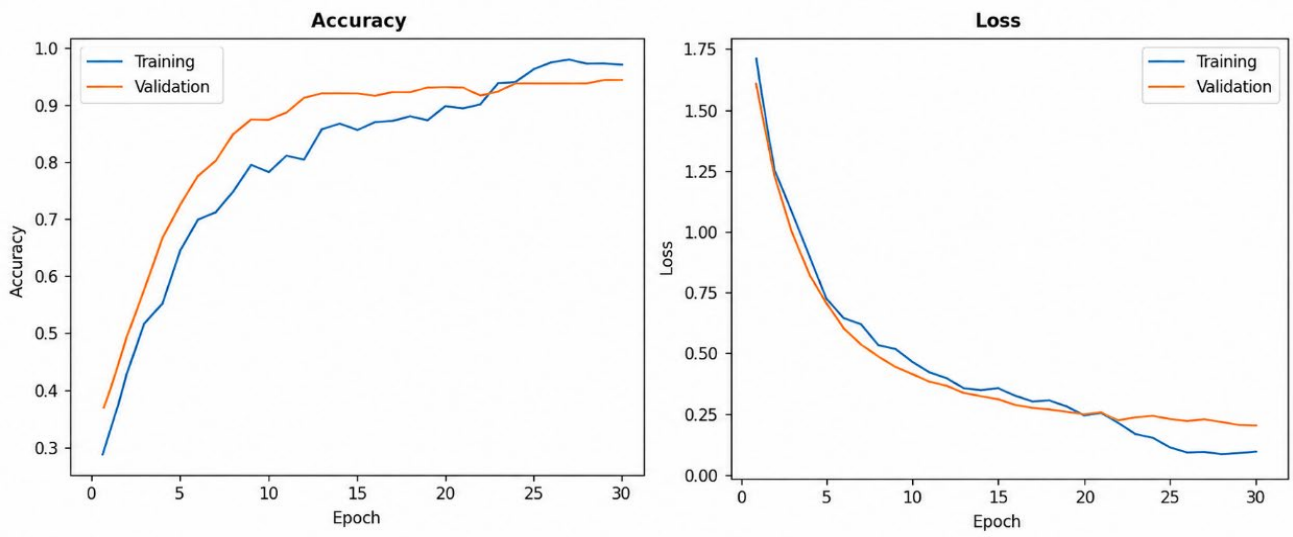


Figure 2. Training and validation accuracy and loss throughout the two-stage training process.

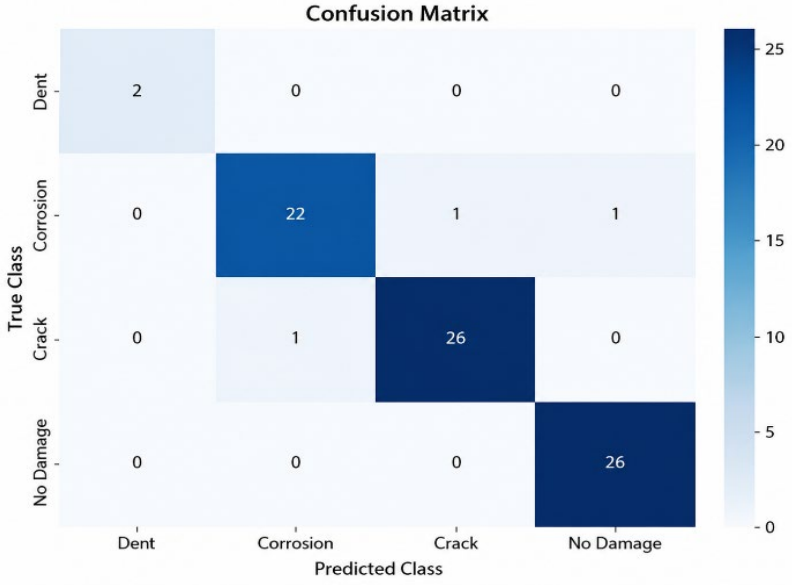


Figure 3. Confusion matrix obtained on the independent test subset (n = 79).

Table 2. Per-class classification performance on the test subset.

Class	Precision	Recall	F1-score	Support
Dent	1.0000	1.0000	1.0000	2
Corrosion	0.9565	0.9167	0.9362	24
Crack	0.9630	0.9630	0.9630	27
No Damage	0.9630	1.0000	0.9811	26

Macro avg.	0.9706	0.9699	0.9701	79
Weighted avg.	0.9619	0.9620	0.9617	79

The independent test subset contained 79 images. As shown in Figure 3, the classifier correctly predicted 76 samples, corresponding to an overall test accuracy of 96.20% and a test loss of 0.1231. The no-damage class achieved perfect recall, while the crack class correctly identified 26 of 27 samples. Corrosion accounted for the largest number of errors, with 22 of 24 samples correctly classified, and confusion was isolated to the crack and no-damage categories. The per-class results summarized in Table show a weighted F1-score of 0.9617 and a macro F1-score of 0.9701. The remaining confusion between corrosion and crack suggests that some samples share visual characteristics such as irregular edges, localized texture changes, and surface discontinuities. Although the overall performance on the independent test subset was strong, the limited number of dent samples should still be considered when interpreting the perfect score obtained for this category.

VI. LIMITATIONS AND FUTURE WORK

The current study represents a preliminary stage in the development of an intelligent aircraft structural inspection framework. Although the proposed classifier demonstrated the ability to learn relevant visual patterns across multiple damage categories, several limitations remain. First, the available dataset size is still limited and does not fully represent the diversity of structural conditions encountered in real operational environments. Variations in illumination, viewing angle, damage severity, and surface materials may affect generalization performance. In addition, the current system performs image-level classification only and does not provide damage localization or dimensional characterization. Future work will focus on expanding the dataset, improving class balance, and incorporating object detection techniques for automatic damage localization. Additional developments will include explainable artificial intelligence methods, geometric measurement modules, and integration with aircraft maintenance documentation to support maintenance decision-making processes.

VII. CONCLUSIONS

This work presented the preliminary development of an aircraft structural damage classification system based on deep learning and transfer learning techniques. A ResNet50 architecture was implemented to classify aircraft images into four structural condition categories: crack, dent, corrosion, and no damage. The obtained results demonstrated the feasibility of applying convolutional neural networks to aircraft structural inspection tasks. They confirmed the proposed methodology's ability to learn relevant visual characteristics associated with different damage types. Although classification errors remain in visually similar categories, the current implementation provides a solid foundation for future improvements. The proposed classifier is the first stage of a broader intelligent inspection framework that incorporates damage localization, geometric characterization, explainable analysis, and maintenance-oriented decision support.

REFERENCES

[1]. Plastropoulos, K. Bardis, G. Yazigi, N. P. Avdelidis, and M. Droznika, "Aircraft skin machine learning-based defect detection and size estimation in visual inspections." Technologies, vol. 12, no. 9, p. 158, 2024.

[2]. Uzun, M. R. Tolun, F. Sari, and F. N. Alpaslan, "Damage detection in aircraft engine borescope inspection using deep learning." Neural Computing and Applications, vol. 37, no. 24, pp. 20105-20126, 2025.

[3]. M. Padsumbiya, V. Brahmbhatt, and S. Thakkar, "Automatic crack detection using a convolutional neural network." Journal of Soft Computing in Civil Engineering, vol. 6, no. 3, pp. 1-17, 2022.

[4]. Brandoli, A. R. de Geus, J. R. Souza, G. Spadon, A. Soares, J. F. Rodrigues Jr., J. Komorowski, and S. Matwin, "Aircraft Fuselage Corrosion Detection Using Artificial Intelligence," Sensors, vol. 21, no. 12, p. 4026, 2021.

[5]. N. Suvittawat, C. Kurniawan, J. Datephanyawat, J. Tay, Z. Liu, D. W. Soh, and N. A. Ribeiro, "Advances in Aircraft Skin Defect Detection Using Computer Vision: A Survey and Comparison of YOLOv9 and RT-DETR Performance," Aerospace, vol. 12, no. 4, p. 356, 2025.

[6]. R. E. Donatus, O. C. Ubadike, M. U. Bonet, S. D. Iyaghiba, I. H. Donatus, and N. I. Mbada, "Automated Aircraft Structural Defect Detection Using Deep Learning and Computer Vision," Mekatronika: Journal of Mechatronics and Intelligent Manufacturing, vol. 7, no. 2, pp. 108-123, 2025.

[7]. Doğru, S. Bouarfa, R. Arizar, and R. Aydoğan, "Using Convolutional Neural Networks to Automate Aircraft Maintenance Visual Inspection," Aerospace, vol. 7, no. 12, 2020.

[8]. Y. Liu, J. Dong, Y. Li, X. Gong, and J. Wang, "A UAV-Based Aircraft Surface Defect Inspection System via External Constraints and Deep Learning," IEEE Transactions on Instrumentation and Measurement, vol. 71, 2022.

[9]. R. Sterkenburg and M. J. Kroes, Aircraft Maintenance and Repair, McGraw-Hill, 2019.

[10]. M. Azimi and G. Pekcan, "Structural health monitoring using extremely compressed data through deep learning," Computer-Aided Civil and Infrastructure Engineering, vol. 35, no. 6, p. 597–614, 2020.

[11]. H. Li, C. Wang, and Y. Liu, "A High-Quality GAN With Data Augmentation for Aircraft Skin Defect Detection Under Limited Data," IEEE Transactions on Instrumentation and Measurement, 2024.

[12]. k. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," Proceedings of CVPR, pp. 770-778, 2016.

[13]. J. Zhao, W. Xie, D. Yu, Q. Yang, S. Meng, and Q. Lyu, "Deep Transfer Learning Approach for Localization of Damage Area in Composite Laminates Using Acoustic Emission Signal," Polymers, vol. 15, no. 6, 2023.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

From Neo4j to Redis: Storing Graph Data and Comparing Graph Algorithm Performance

Anikó Vágner
Department of Information Technology
Faculty of Informatics
University of Debrecen
Debrecen, Hungary
vagner.aniko@inf.unideb.hu
ORCID:0000-0001-8843-2403

Fanni Belics
Department of Information Technology
Faculty of Informatics
University of Debrecen
Debrecen, Hungary
belics.fanni@inf.unideb.hu
ORCID:0009-0003-2042-2817

Cristina Maria Costa Rodrigues
Department of Computer Science and Quantitative Methods
School of Management and Technology
Polytechnic Institute of Santarém
Santarém, Portugal
cristina.rodrigues@esg.ipsantarem.pt
ORCID: 0000-0001-7442-4277

Abstract
Graph databases provide efficient storage and traversal mechanisms for highly connected data, whereas Redis offers high-performance in-memory data management through general-purpose data structures. Although Redis is widely adopted in modern applications, it does not natively support the property graph model. This paper presents a formal mapping of the Neo4j property graph model onto native Redis data structures, including sets, sorted sets, and hashes. The proposed representation preserves node identifiers, relationship identifiers, relationship directions, node labels, relationship types, and properties without relying on graph-specific Redis extensions. Based on the proposed model, several graph algorithms, including breadth-first search, shortest path, all-shortest-paths, weakly connected components, and degree centrality, were implemented and evaluated. Performance measurements were conducted using the Northwind and Pokec Social Network datasets and compared with the native implementations provided by Neo4j. The experimental results show that Neo4j consistently achieves superior performance, particularly for traversal-intensive algorithms executed on large graphs. Nevertheless, the proposed Redis representation demonstrates that property graphs can be modeled effectively using native Redis data structures, offering a practical alternative for applications that already use Redis as their primary data store and require only moderate graph-processing capabilities.

Keywords*:* Neo4j, Redis, property graph, graph database, graph transformation, graph algorithms, NoSQL

I. INTRODUCTION

In the NoSQL era, it is hard to choose an appropriate database. Firstly, there are four main categories: key-value, document, column-family, and graph. Then the categories are not clean, and each category can contain many databases with different data structures and data processing capabilities. [1] Graph data structures are important in modern computing due to the growing need to model complex relationships in domains such as social networks, recommendation systems, fraud detection, and knowledge graphs. Among many graph databases, Neo4j is the first in the database engines ranking (https://db-engines.com/en/ranking). [2] Neo4j implements the property graph model, where nodes and relationships may contain arbitrary properties. The database supports the declarative Cipher query language, enabling developers to express complex graph operations concisely and readably. Neo4j also provides optimized traversal mechanisms and built-in support for graph analytics. [3] In parallel, in-memory key-value stores such as Redis have gained popularity due to their high performance and low-latency data access. Redis is a popular in-memory database that supports various native data structures, including strings, hashes, lists, sets, sorted sets, streams, and bitmaps. Although Redis is not primarily designed as a graph database, graph structures can be represented efficiently using combinations of these data structures. Such representations may offer significant performance advantages for specific graph workloads due to Redis's

memory-centric architecture. However, unlike native graph databases, Redis does not provide built-in support for graph traversal operations, raising the question of how efficiently graph algorithms can be executed on such representations. [4], [5]. This study investigates the performance differences between a native graph database and Redis-based graph implementations. Specifically, we compare Neo4j with a Redis graph modeling approach under identical workloads. The evaluation focuses on classical graph algorithms, including breadth-first search and shortest path computation.

II. RELATED WORK

1. Graph database models and foundations

The theoretical foundations of graph data models are well established. Angles and Gutierrez [6] provide a comprehensive survey of graph database models, outlining early approaches and the evolution toward modern property graph systems. The property graph model, which allows both nodes and relationships to carry arbitrary key-value properties, has become the dominant paradigm in practical systems. A widely adopted implementation of this model is Neo4j, which is described in detail in the foundational book by Robinson, Webber, and Eifrem [7]. Neo4j introduced the Cipher query language and popularized declarative graph querying for property graphs. The conceptual design of Neo4j and its property graph model is further documented in an industry white paper [8], which emphasizes traversal efficiency and flexible schema design. Query language design for graph systems has also been an important research topic. Cipher, as described in research papers [9], [10], enables expressive pattern-matching queries over graph structures while abstracting low-level traversal logic. This has contributed significantly to the adoption of graph databases in real-world systems.

2. Benchmarking and performance evaluation of graph databases

Several studies have investigated the performance characteristics of graph database systems. The Linked Data Benchmark Council (LDBC) Social Network Benchmark (SNB) provides a standardized workload for evaluating graph database performance under realistic social network scenarios [11]. Rusu and Huang [12] present an in-depth analysis of graph database systems using LDBC SNB, highlighting differences in query execution efficiency and system scalability. More recent comparative studies have evaluated multiple graph database platforms under identical workloads. Monteiro et al. [13] compare Neo4j, JanusGraph, NebulaGraph, and TigerGraph, focusing on query execution time, data loading performance, and resource consumption. Their results show that performance varies significantly depending on workload characteristics and system architecture. Sandell et al. [14] extend this line of work by comparing Neo4j with relational and multi-model databases, including ArangoDB and MySQL. Their findings indicate that graph databases generally outperform relational systems for traversal-heavy queries, while performance differences depend on query complexity and data distribution. In addition to system-level benchmarks, recent research has explored optimization techniques for improving graph query execution. Baumstark et al. [15] propose adaptive query compilation techniques that reduce runtime overhead and improve execution efficiency for graph queries, particularly under mixed workloads. A more recent study by Lupu et al. [16] evaluates concurrent access patterns in relational and graph NoSQL databases, showing that graph databases perform efficiently under traversal-intensive workloads, although concurrency introduces measurable performance constraints.

3. Redis and in-memory data systems

In parallel with graph database systems, in-memory data stores have gained significant attention for their low-latency performance. Redis, described in its official documentation [4] and in Sanfilippo's foundational design

notes [17], is a high-performance in-memory data structure store that supports key-value pairs, sets, hashes, and other primitives. Redis has been widely studied as a general-purpose in-memory system. Almeida et al. [18] compare Redis with Memcached, MySQL, and PostgreSQL using the YCSB benchmark, demonstrating that Redis achieves high throughput and low latency, particularly in read-intensive workloads. Redis is also well documented in system-level literature, including practitioner-oriented texts such as Redis in Action [5] and The Little Redis Book [19]. Although Redis is not a native graph database, graph structures can be modeled. This has led to the development of RedisGraph, a graph database module built on Redis that leverages optimized in-memory representations and GraphBLAS-based computations [20]. RedisGraph was excluded from the evaluation because it is a legacy component with limited active development and is therefore not considered representative of modern graph database systems.

4. NoSQL data models and system comparisons

Several studies have examined NoSQL databases from a model-level perspective. Kamal et al. [21] provide a qualitative comparison of Redis, Neo4j, and MongoDB, focusing on structural differences between key-value, document, and graph models. Their work highlights the conceptual mapping between different NoSQL paradigms but does not evaluate runtime performance.

III. MATHEMATICAL REPRESENTATION OF THE NEO4J PROPERTY GRAPH MODEL

A Neo4j database follows the labeled property graph model, in which data is represented as a directed graph of nodes and relationships, enriched with labels and key–value properties. The following formalization is based on the property graph model described by Angles and Gutierrez, as well as the latter survey of property graph database systems [22], [23]. We changed the definition because we want to store the graph in Redis, where the node and relationship identifiers will be important. Let L be a countable set of node labels, and T be a countable set of relationship types. A property graph is defined as a tuple G = (N, R, S, D, PN, PR, B, Q), where N is a finite set of node identifiers, R is a finite set of relationship instances, each identified by a unique relationship identifier. S: R- >N maps each relationship identifier to its source node identifier, D: R->N maps each relationship identifier to its destination node identifier, PN: N×K->V is a finite partial function mapping a node identifier and a property key to a value; similarly PR: R×K->V is a finite partial function mapping a relationship identifier and a property key to a value; B: N->2L maps each node identifier to a finite (possibly empty) set of labels (where 2L is the power set of L allowing a node to have zero, one, or multiple labels); and Q: R->T maps each relationship identifier to a relationship type. This definition corresponds to the Neo4j data model, where relationships are directed and have a single relationship type [3]. A node can be represented as: n = (n_id, B(n), PN(n)) where n_id is the node identifier, B(n_id) is the set of labels assigned to the node, and PN(n_id) contains its optional key-value properties. A relationship can therefore be represented as r = (r_id, S(r_id), D(r_id), Q(r_id), PR(r_id)) where S(r_id) and D(r_id) define the direction of the relationship with its source and destination, Q(r_id) defines its type, and PR(r_id) stores optional properties.

IV. NEO4J PROPERTY GRAPH MODEL REALIZATION IN REDIS

To evaluate graph algorithm performance independently from the database engine, the Neo4j property graph model is transformed into a Redis-based graph representation. The transformation preserves the graph semantics, including node identities, relationship directions, relationship types, labels, and properties.

The node identifiers will be stored in a set named nodes, each in the form n_id, where id is a positive integer. The form of the relationship identifier will be similar to r_id, where id is a positive integer. The relationship does

not exist in isolation; we store its source and destination nodes. The sources in the relationship will be stored in a zset (sorted set), whose name is r_sources, and the key is the r_id, the score is the id from the n_id, because the score should be a floating-point number, so we store the integer part of the id, and pay attention to the conversion when we use it. We choose the zset because it is sorted by score, so searching for the node identifier is faster. The destination is also stored in a zset; similarly, the zset name is r_destinations. The properties of the nodes are stored in hashes; each node has a hash named n_id. Similarly, the properties of the relationships and the names of the hashes are r_id. Since a relationship can have only one type, r_types named can store the r_id->type values. And finally, the node labels are stored in sets; each node has its own set named labels_n_id, which contains the labels. A formalized form of the Redis data structures for the Neo4j graph property is

$RG=(S_{nodes}, Z_{r_sources}, Z_{r_destinations}, \{H_n| n\in S_{nodes}\},\{H_r|r\in R, H_{r_types}, \{S_{label,n}|n\in S_{nodes}\})$, where R is the set of r_ids. S is a set, Z is a sorted set (zset), and H is a Hash.

V. TRANSFORMATION OF THE PROPERTY GRAPH MODEL INTO REDIS DATA STRUCTURES

To evaluate graph algorithms using Redis, the previously defined property graph model is transformed into a set of native Redis data structures. The transformation preserves the semantics of the original graph, including node and relationship identities, relationship directions, relationship types, labels, and properties.

The transformation can be defined as a mapping M: G -> RG, where G is the property graph representation, and RG is the corresponding Redis representation.

The node identifier set is mapped to a Redis set:
$M(N)=S_{nodes}$, where: $S_{nodes}=\{n_{id}| n\in N\}$
Each element of S_{nodes} represents a unique node identifier.

Relationship identifiers are stored in the source and destination ZSet-s of Redis.

Each relationship is represented by its identifier because relationships are first-class elements of the property graph model.

Two separate mappings preserve the direction of relationships:
$M(S)=Z_{sources}$ and $M(D)=Z_{destinations}$
where each Redis sorted set stores a mapping between a relationship identifier and the corresponding node identifier:
$Z_{sources}[r_id]=S(r_id)$ and $Z_{destinations}[r_id]=D(r_id)$.
$M(R)=Z_{sources}[r_id] \cup Z_{destinations}[r_id]$

The relationship type mapping is stored as a Redis hash:
$M(Q)=H_{types}$, where: $H_{types}[r_id]=Q(r_id)$.

Because each relationship has exactly one type in Neo4j, a single value is stored for every relationship identifier.

Node properties are represented using Redis hashes. For every node identifier (n_id), a separate hash is created:
$M(PN(n_id))=H_{n_id}$, where: $H_{n_id}[k]=PN(n_id,k)$
Relationship properties are stored analogously:
$M(PR(r_id))=H_{r_id}$, where: $H_{r_id}[k]=PR(r_id,k)$

Redis sets represent node labels:

$M(B(n_id))=S_{labels,n_id}$, where: $S_{labels,n_id}=B(n_id)$

For both Neo4j and Redis-based graph representations, the storage requirement depends on the number of nodes, the number of edges, and the amount of associated metadata, including labels and properties assigned to nodes and relationships.

VI. THE DATASET AND QUERIES

Our first dataset is Northwind (https://github.com/neo4j-graph-examples/northwind). To create it, we downloaded the Cipher statements and executed them. It has about 1000 nodes and 3000 relationships. Our second dataset is the Pokec Social Network (SNAP), which can be downloaded from Stanford University's website (https://snap.stanford.edu/data/soc-Pokec.html). We used only the relationships file, in which the node_ids and their relationships are such that the source_node_id and destination_node_id of the relationship are present. For our queries, this information is enough. It has about 1.6 million nodes and 33 million relationships.

We executed 4 main queries, a few of which had parameters; we varied the node_ids or the maximum path length. The four queries are the following:

- Q1 (Breadth-First Traversal - BFS): Starting from a given node_id, retrieve all distinct nodes reachable by following directed relationships with a maximum path length and return their count. This query simulates a breadth-first exploration of the graph up to the third level of connectivity. We executed this query at 30 levels of connectivity as well.
- Q2 (Shortest Path): Given two node_ids, determine the shortest path between them with a maximum 10 path length relationships. This query evaluates the graph database's ability to perform path-finding operations across heterogeneous node and relationship types.
- Q3 (All Shortest Paths): Given two node_ids, determine all the shortest paths between them with a maximum path length. The maximum length paths were 2, 5, 6, 10, and 15.
- Q4 (Connected Components): The Weakly Connected Components algorithm was executed on the full graph to identify disconnected subgraphs. The method assigns a component identifier to each node, grouping nodes that are reachable from each other when edge directions are ignored. This workload evaluates the graph's global connectivity structure and the efficiency of component labeling in heterogeneous graphs.
- Q5 (Degree Centrality): The degree centrality algorithm was executed on the full graph to measure the number of direct relationships associated with each node. The resulting ranking identifies the most connected entities, providing insights into structural importance based on immediate connectivity.

Queries 4 and 5 need to build the in-memory graph with gds. graph.project statement before we execute the queries in Neo4j. The in-memory graph building is not included in the performance comparison. Neo4j's built-in algorithms do not work without an in-memory graph.

VII. ENVIRONMENT

The host computer is a Lenovo laptop with an Intel i5 1GHz processor and 8GB RAM; the operating system is Microsoft Windows 11 Pro, on which we installed Oracle VirtualBox 7.2.6, on which Ubuntu server (26.0.4

LTS) was installed twice, one for the Neo4j (2026.05.0) server and the other for the Redis server (8.0.5). Each virtual machine gets 2 CPUs, 2GB of memory, and 100 GB of disk space; however, we did not use all of them. The host computer was equipped with PyCharm (2026.1.2), on which Python (3.14.6) code was executed; the databases required drivers: the Redis module (8.0.0) named Redis and the Neo4j module (6.2.0) named Neo4j. The second, Pokec Social Network dataset, was too large, and Redis needed more memory since it is a memory database; it requested 12GB of memory. So for this dataset, we had to change the environment; the Lenovo laptop executed the Python code, but the virtual boxes moved to two other, but equal, computers, namely the Dell Pro Slim desktop computers with Windows 11 Enterprise Edition operating system; the processors are i5 with 2.6Ghz, and they have 16GB RAM. The VirtualBox version is 7.0.26, and both virtual machines have 12GB RAM. Since we wanted to see where Neo4j throws out-of-memory exceptions, we also executed the queries with 4GB and 2GB of RAM.

VIII. PERFORMANCE OPTIMIZATION IN REDIS

While writing the queries, it turned out that the zset for the relationships is inefficient, because after we get the relationship ID, we have to ask Redis again for the node ID. Because of this, we built two additional data structures for the relations: one for the sources and another for the destinations. Each of them is a hash set, where the key is neighbor_source_n_id and neighbor_destination_n_id; the key is the source or destination node id, and the value is the relationship id. In NoSQL systems, data models are typically designed based on query patterns rather than normalization rules, often intentionally introducing redundancy to optimize read performance and query efficiency. These two data structures are used in all the executed algorithms.

IX. RESULTS

We executed our queries 20 times on the databases and calculated an average execution time for each query. Table 1 shows the results of the execution on the Northwind database, while Table 2 shows the results of the execution on the Pokec Social Network database. The ML shows the maximum-length parameter for each algorithm. The node_ids were chosen so that the maximum length can be reached between them.

Table 1. Execution times of queries on Northwind

	Execution times (in seconds)	
Queries	Neo4j	Redis
Q1 (ml=3)	0.0578	4.000
Q1 (ml=30)	0.0083	2.000
Q2 (ml=6)	0.0154	2.000
Q3 (ml=2)	0.0337	0.0412
Q3 (ml=5)	0.0215	4.6840
Q3 (ml=6)	0.0293	12.7343
Q3 (ml=10)	0.0227	7.0952
Q3 (ml=15)	0.0342	8.5982
Q4	0.7168	8.000
Q5	0.0549	6.000

Table 2.Execution times of queries on the Pokec Social Network

	Execution times (in seconds)

Queries	Neo4j 12GB RAM	Neo4j 4GB RAM	Neo4j 2GB RAM	Redis
Q1 (ml=3)	0.0376	0.0380	0.0625	0.8535
Q1 (ml=30)	5.4890	6.1358	8.9395	25452.1245
Q2 (ml=6)	0.0129	0.0290	0.0242	2.4648
Q3 (ml=2)	0.0145	0.0120	0.0111	0.0113
Q3 (ml=5)	0.0311	0.0159	0.0191	2144.6362
Q3 (ml=6)	0.0163	0.0204	0.0227	3175.0377
Q3 (ml=10)	0.0068	0.0095	0.0115	3889.5314
Q3 (ml=15)	0.0058	0.0083	0.0103	3883.2865
Q4	89.8378	85.5693	OutOfMemory	5337.1493
Q5	0.4466	0.4486	1.0999	4838.9106

The Northwind dataset is a relatively small graph that enables comparison of graph traversal algorithms under a moderate workload. As shown in Table 1, Neo4j consistently outperformed the Redis implementation for almost all evaluated queries. The difference is particularly noticeable for traversal-intensive algorithms such as breadth-first search (Q1), shortest path (Q2), weakly connected components (Q4), and degree centrality (Q5), where Redis took several seconds, whereas Neo4j completed the same operations within a fraction of a second. An exception can be observed for the shortest-path query (Q3), which has a maximum path length of 2. In this case, both systems achieved nearly identical execution times. Since only a very limited neighborhood has to be explored, the traversal overhead remains negligible in both implementations. However, as the maximum path length increases, the execution time of the Redis implementation grows rapidly, while Neo4j remains almost unaffected. This behavior demonstrates the efficiency of Neo4j's native graph traversal engine compared to application-level graph traversal implemented on top of general-purpose Redis data structures. The second dataset, the Pokec Social Network dataset, represents a significantly larger graph and therefore provides a more realistic benchmark for graph analytics. Neo4j consistently achieved low execution times for all evaluated algorithms, even when the available memory was reduced from 12 GB to 4 GB and further to 2 GB. Although slight performance degradation is observed for some queries, overall execution times remain relatively stable, indicating efficient memory management and optimized graph storage. The Redis implementation exhibits similar behavior to Neo4j only for the shortest-path query with a maximum path length of 2, where both systems produced almost identical execution times. For larger traversal depths, however, the execution time increased dramatically. For example, the execution time of Q3 increased from 0.0113 s for a maximum path length of two to more than 3,800 s for path lengths of ten and fifteen. Similar behavior can be observed for weakly connected components (Q4) and degree centrality (Q5), where the Redis implementation required more than one hour to complete the computations. An additional observation is that Neo4j successfully executed all benchmark queries with both 12 GB and 4 GB of available memory. Even with only 2 GB of memory, most algorithms completed successfully, although the weakly connected components algorithm could not be executed because of an out-of-memory condition. This result indicates that Neo4j's native graph storage and traversal mechanisms remain efficient even under memory constraints. The observed performance differences can primarily be attributed to the underlying data models. Neo4j stores graph structures natively and executes traversal algorithms directly within the database engine. In contrast, the proposed Redis representation relies exclusively on native Redis data structures, including sets, sorted sets, and hashes. Consequently, graph traversal must be performed explicitly by the application, requiring multiple Redis operations for each traversal step. While this approach introduces considerable execution overhead for traversal-intensive algorithms, it demonstrates that graph structures can be represented using standard Redis data types without requiring specialized graph database extensions. Overall, the experimental results show that Neo4j provides substantially better performance for graph analytics and complex graph traversals. Nevertheless, the proposed Redis representation offers a flexible and portable solution

for applications that already employ Redis as their primary data store and require only occasional graph processing without introducing an additional graph database management system.

X. FUTURE WORK

Since Redis supports clustering and data partitioning, future work will investigate distributed graph processing techniques that exploit multiple Redis nodes. The proposed representation relies on a limited number of global indexes for relationship lookup. Additional indexing strategies, such as neighborhood indexes, path indexes, or application-specific auxiliary structures, could further reduce graph traversal costs and improve the performance of complex graph algorithms. Another promising direction is evaluating the proposed model using additional graph algorithms, including community detection, PageRank, betweenness centrality, and graph pattern matching. The proposed mapping can serve as the foundation for an automated graph migration framework that transforms Neo4j property graphs into Redis-native representations while preserving the graph structure, labels, relationship types, and properties. Such a framework would simplify the integration of graph-based applications into existing Redis infrastructures. Another important research direction is the formal verification of the proposed transformation to prove that it is information-preserving and reversible under well-defined constraints. Finally, while the proposed mapping fully supports node labels, relationship types, and properties, the experimental evaluation presented in this paper concentrates primarily on graph structure and traversal performance. A more detailed investigation of label- and property-intensive workloads, including filtering, indexing, and property-based graph queries, remains an important direction for future research.

XI. CONCLUSION

This paper presented a Redis-based representation of the Neo4j property graph model using only native Redis data structures, including sets, sorted sets, and hashes. A formal mapping was proposed that preserves the fundamental components of the property graph model, including node identifiers, relationship identifiers, relationship directions, node labels, relationship types, and properties. Based on this representation, several graph algorithms were implemented to evaluate the feasibility of executing graph operations without relying on dedicated graph database extensions. The experimental evaluation conducted on the Northwind and Pokec Social Network datasets demonstrated that Neo4j consistently outperformed the Redis implementation, particularly for traversal-intensive algorithms and large-scale graphs. The results indicate that native graph storage and optimized traversal mechanisms provide a significant advantage over graph processing implemented on top of general-purpose key-value data structures. Nevertheless, the proposed Redis representation successfully preserves the semantics of the original property graph while providing a flexible solution for applications that already utilize Redis as their primary data store. Although the Redis implementation cannot compete with native graph databases in traversal performance, it offers an alternative for integrating graph-based functionality into existing Redis-based systems without introducing additional database technologies. Furthermore, the proposed formal mapping establishes a foundation for future research on graph data migration, optimization techniques, distributed graph processing, and automatic transformation between graph and key-value database models. Overall, the proposed model demonstrates that native Redis data structures can represent the complete Neo4j property graph model while preserving its structural semantics, providing a basis for further research into efficient graph processing in key-value databases.

REFERENCES

[1]. P. J. Sadalage and M. Fowler, NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Upper Saddle River, NJ, USA: Addison-Wesley Professional, 2012.



[2]. M. Kleppmann, Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems. Sebastopol, CA, USA: O'Reilly Media, 2017.
[3]. Neo4j Inc., Graph Database Concepts: Nodes, Relationships, Labels, and Relationship Types. Available: https://neo4j.com/docs/. Accessed: 7 Jul. 2026.
[4]. Redis Inc., Redis Documentation. Available: https://redis.io/docs/. Accessed: 7 Jul. 2026.
[5]. J. Carlson, Redis in Action. Shelter Island, NY, USA: Manning Publications, 2013.
[6]. R. Angles and C. Gutierrez, "Survey of Graph Database Models," ACM Computing Surveys, vol. 40, no. 1, pp. 1–39, 2008.
[7]. I. Robinson, J. Webber, and E. Eifrem, Graph Databases: New Opportunities for Connected Data. Sebastopol, CA, USA: O'Reilly Media, 2015.
[8]. E. Eifrem, "The Property Graph Database Model," Neo4j White Paper, Neo4j Inc., 2013.
[9]. F. Holzschuher and R. Peinl, "Querying a Graph Database – Language Selection and Performance Considerations," *Journal of Computer and System Sciences*, vol. 82, no. 1, pp. 45–68, 2016.
[10]. N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, and A. Taylor, "Cipher: An Evolving Query Language for Property Graphs," in Proceedings of the 2018 International Conference on Management of Data (SIGMOD), pp. 1433-1445, 2018.
[11]. Linked Data Benchmark Council, Social Network Benchmark (SNB) Specification, LDBC, 2019. Available: https://ldbcouncil.org/benchmarks/snb/. Accessed: 7 Jul. 2026.
[12]. F. Rusu and Z. Huang, "In-Depth Benchmarking of Graph Database Systems with the Linked Data Benchmark Council Social Network Benchmark," arXiv preprint arXiv:1907.07405, 2019.
[13]. J. Monteiro, F. Sá, and J. Bernardino, "Experimental Evaluation of Graph Databases: JanusGraph, Nebula Graph, Neo4j, and TigerGraph," Applied Sciences, vol. 13, no. 9, p. 5770, 2023.
[14]. J. Sandell, E. Asplund, W. Y. Ayele, and M. Duneld, "Performance Comparison Analysis of ArangoDB, MySQL, and Neo4j," arXiv preprint arXiv:2401.17482, 2024.
[15]. A. Baumstark, M. A. Jibril, and K.-U. Sattler, "Adaptive Query Compilation in Graph Databases," Distributed and Parallel Databases, vol. 41, pp. 359–386, 2023.
[16]. E. Lupu, A. Olteanu, and A. D. Ioniță, "Concurrent Access Performance Comparison Between Relational Databases and Graph NoSQL Databases for Complex Algorithms," Applied Sciences, vol. 14, no. 21, p. 9867, 2024.
[17]. S. Sanfilippo, Redis Design Notes and Documentation. Redis Inc. Available: https://redis.io. Accessed: 7 Jul. 2026.
[18]. D. Almeida, M. Lopes, L. Saraiva, M. Abbasi, P. Martins, J. Silva, and P. Váz, "Performance Comparison of Redis, Memcached, MySQL, and PostgreSQL," in Proceedings of IEEE SmartTech-IC, 2023, pp. 902–907.
[19]. K. Seguin, The Little Redis Book. 2018.
[20]. O. Rebeiro-Hargrave, J. Nider, A. Zaks, and M. Opher, "RedisGraph: A Graph Database Based on Redis," in Proceedings of IEEE ICDE Workshops, 2019, pp. 1–8.
[21]. S. H. Kamal, H. H. Elazhary, and E. E. Hassanein, "A Qualitative Comparison of NoSQL Data Stores," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 2, pp. 330–338, 2019.
[22]. R. Angles, "The Property Graph Database Model," in Proceedings of 12th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW), 2018.
[23]. R. Angles, M. Arenas, P. Barceló, A. Hogan, J. Reutter, and D. Vrgoč, "Foundations of Modern Query Languages for Graph Databases," ACM Computing Surveys, vol. 50, no. 5, Art. no. 68, 2018.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Open Cloud Voice Service Architecture

Alex Macia Fiteni
Dept of Computer Science and Technology
University of Alicante
San Vicente del Raspeig, Spain
amf107@alu.ua.es
ORCID: 0000-0001-8462-9428

Iren Lorenzo Fonseca
Dept of Computer Science and Technology
University of Alicante
San Vicente del Raspeig, Spain
iren.fonseca@ua.es
ORCID: 0000-0003-3597-4836

Jaime E. Chiang Cruz
Hydraulic Research Center
Technological University of La Habana "José Antonio Echeverría"
La Habana, Cuba
jaimeerch@civil.edu.cu
ORCID: 0009-0001-4899-701X

Francisco Maciá Pérez
Dept of Computer Science and Technology
University of Alicante
San Vicente del Raspeig, Spain
pmacia@dtic.ua.es
ORCID: 0000-0002-2516-4728

Abstract
We are facing an increasingly aging population, with a growing demand for health and care services. Technology has proved highly useful in this field, but proposed solutions must be adapted as closely as possible to the characteristics and needs of older adults. Voice-based interfaces, or Voice User Interfaces (VUIs), offer many advantages in this regard. Still, it remains crucial that they be based on voice and conversational models specifically designed for older adults. Likewise, this type of interface relies on a sophisticated cloud service infrastructure that must be versatile and secure while remaining affordable for any project. This paper proposes, from an architectural perspective, a cloud-based voice service, SAM Voice Service, that enables the design of VUI-based devices for older adults and people with disabilities. As a case study, it presents a companionship service for older adults and people with care needs, developed jointly with the social services departments of two local councils. More specifically, it presents the design of a physical companionship device for older adults based on VUI and natural language, together with a deployment architecture for the cloud-based voice service and a development scenario for all service components.

Keywords: Voice User Interface (VUI), Natural Language Processing (NLP), Cloud Service, Distributed Systems, Elderly, Active and Assisted Living (AAL)

I. INTRODUCTION

Population aging is increasing the number and proportion of people aged 65 years and over, placing growing pressure on health and care services [1-4]. Technology-based services have been proposed to address some of these needs, with support from initiatives such as the Ambient Assisted Living and Active and Assisted Living programs [3], [5-8]. However, as solutions to this new situation emerge, new problems are also being identified. One such problem is that most solutions rely on user interfaces not specifically designed for older adults [9], [10], or, at best, are adaptations of existing interfaces. It is essential to consider the specific and sensitive characteristics of older adults when developing devices and interfaces for them [11], [12]. This paper focuses on Voice User Interfaces (VUIs) supported by voice and dialogue models tailored to older adults [11], [13]. Existing studies suggest that older users can accept voice-based technologies when they provide meaningful support [14-16]. Nevertheless, delivering such services requires an architecture that integrates specialized voice and conversational components while accommodating differing technical and operational requirements [17], [18].



Accordingly, this article proposes a set of architectures for a cloud-based voice service, spanning conceptual, technical, and deployment levels. The framework is intended to support realistic service scenarios adapted to the characteristics, available resources, and budgets of different projects and communities. By accommodating diverse technologies, providers, and relevant open-source initiatives, it aims to enable scalable, reliable, and economically and technically sustainable deployments. The contribution of this work is not a new speech or dialogue algorithm, but an open, modular cloud architecture that organizes voice-processing components into independently deployable services adapted to older adults, social-care scenarios, and provider-controlled deployment. The remainder of the paper is organized as follows. Section 2 reviews the relevant background and state of the art. Section 3 presents the problem definition and proposed architecture, while Section 4 describes its technical implementation. Section 5 presents the companionship-service case study, and Section 6 concludes the paper. All project materials and source code have been made available to the community through the public repositories detailed in the paper.

II. STATE OF THE ART

Europe's population is aging: at the beginning of 2025, people aged 65 years and over accounted for 22% of the EU population (99 million), including 6% aged 80 years and over [1]. Eurostat projects that, despite a decline in the total population from 451.8 million in 2025 to 445.0 million in 2050, the number of working-age people per person aged over 65 will fall from more than three to fewer than two [1], [2]. This trend is likely to increase demand for health and care services, as 60% of EU residents aged over 65 reported at least one chronic condition and 52% reported activity limitations in 2023 [4]. Accordingly, care and companionship for older adults are increasingly important areas for research and service development [3]. The main objective of our research is to design Voice User Interfaces (VUIs) for services intended for older adults and people with disabilities. These VUIs must be user-friendly and adapted as closely as possible to their characteristics and needs [11], [12]. The development of commercial voice assistants, alongside growing academic interest, indicates that voice-based interaction has reached a sufficient level of maturity to support services for older adults and people with disabilities [9], [10], [18], [19]. Although commercial platforms perform very well and provide a great many facilities for developing projects that wish to incorporate NLP, even free of charge, in practice, they still present many restrictions [20], [21]. From the perspective of services for older adults, the main restrictions arise from the fact that voice and conversational models are neither specifically designed nor adapted to the characteristics and needs of older adults. They do not facilitate integration with other systems and IoT devices, and there are concerns about information privacy, since conversations must pass through third-party servers, where they are stored to improve and refine dialogue and voice models [11], [13]. It is essential that services and devices designed for older adults explicitly account for the highly specific and defining characteristics, constraints, and idiosyncrasies of this sensitive population segment [11], [12], [22], [23]. Although older adults may initially perceive assistive technologies as replacing family members or trusted carers, studies suggest that voice-based systems can be well accepted when they provide meaningful support [14-16]. Many of the proposals reviewed are based on cloud applications, with mobile-device interfaces that connect social workers, doctors, and, in some cases, family members to the care of older adults. However, in most cases, these proposals do not involve the older person directly or offer them support that makes them feel genuinely accompanied [24], [25], [12].

III. PROBLEM DEFINITION AND PROPOSED APPROACH

The problem addressed in this work is the cost of developing and maintaining services for older adults that provide user interfaces tailored to their needs. To address this challenge, we propose the SAM Voice Service. This cloud-based platform supports SAM Devices: virtual assistants intended to assist and accompany older adults at home through voice-based interaction. Given the multiple stakeholders, technologies, and deployment scenarios involved, this paper presents an integrated architectural view of SAM Voice Service at conceptual, technical, and deployment levels. The proposed service and microservice-based architecture (Figure 1) is

designed to integrate technologies from different providers and to adapt computational resources as service requirements and available technologies evolve.

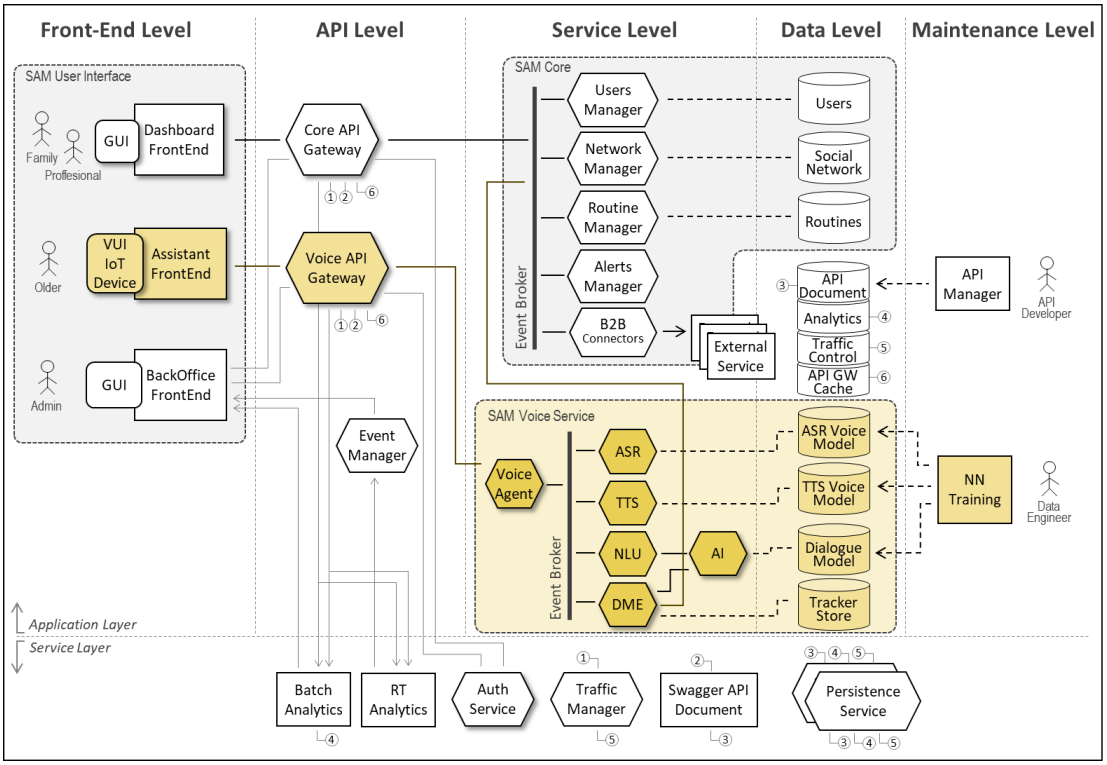


Figure 1. SAM Service high-level conceptual architecture.

For the high-level architecture (Figure 1), an n-tier architectural style has been used, structuring the elements into levels and layers. SAM Voice Service is located at the service and data levels, namely the Service Level and Data Level. Its purpose is to support the voice assistant by performing artificial intelligence (AI) tasks that cannot feasibly be performed on the IoT device or the user-interface device, either because of their high computational cost or because of the business logic they involve. The elements of the architecture are defined primarily as services, following an architectural pattern based on Service-Oriented Architectures (SOA) and microservices [26], [27]. An effort has been made to strike a balance in the number of services specified, so that they are as functionally independent as possible, without generating so many microservices that future maintenance would become complex. The Voice Agent orchestrates the proposed services. Figure 2 shows the activity diagram for the orchestration carried out by this service. The flow begins when the Voice Agent receives an audio request, typically from the Voice API Gateway. The Voice Agent invokes the Automatic Speech Recognition (ASR), also known as Speech-to-Text (STT), service to convert audio into a text request. This request is then processed by the Natural Language Understanding service, which extracts the relevant intents and entities using the Dialogue Model.

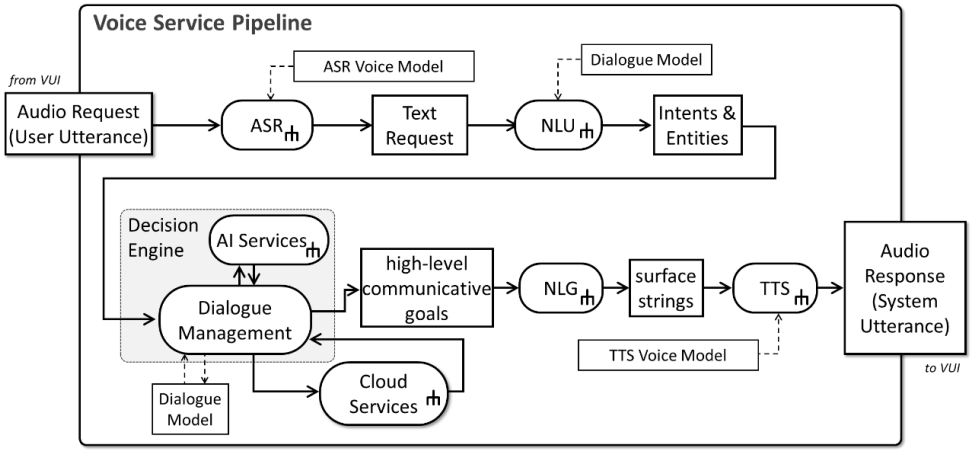


Figure 2. SAM Voice Service Activity Diagram.

The Voice Agent forwards these results to the Dialogue Management Engine, which determines the appropriate response and, where necessary, invokes external services in SAM Core through the SAM Core Event Broker. The resulting response is passed to the Natural Language Generation module and then to the Text-to-Speech service, which produces an audio response. Finally, the Voice Agent returns the response to the Voice API Gateway.

IV. TECHNICAL ARCHITECTURE

Figure 3 summarises the technical architecture of the SAM Voice Service and its relationship with the SAM Device, the SAM Core, and external AI services. The implementation uses Mozilla DeepSpeech for ASR, Rasa for NLU and dialogue management, and Mozilla TTS for speech synthesis. These components are deployed as separate services so that speech recognition, dialogue handling, and speech synthesis can be maintained, replaced, or scaled independently.

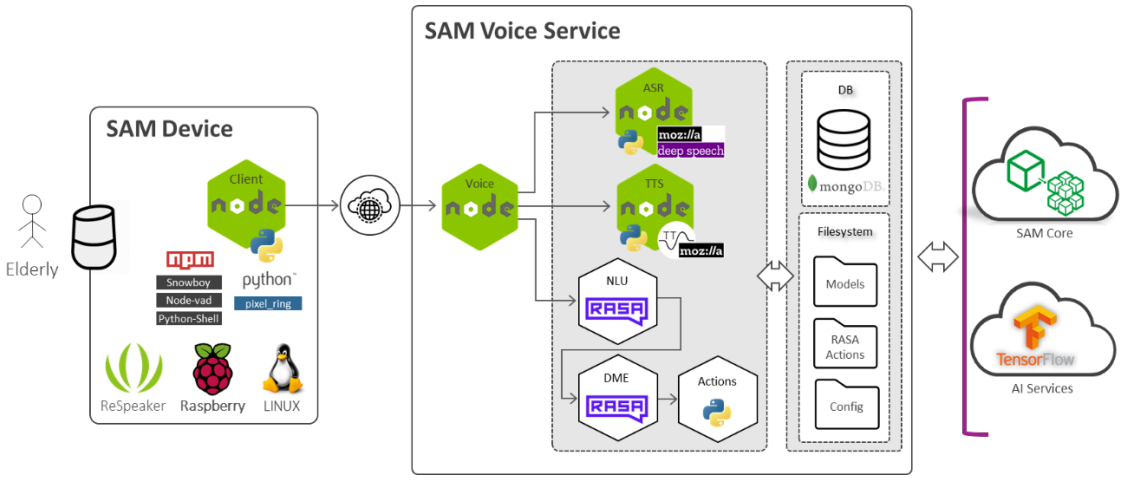


Figure 3. Conceptual technical architecture of the voice service (SAM Voice Service).

Recognition and dialogue performance depend on how well the corresponding models represent the target population and service domain. For this reason, the architecture supports customized ASR and conversational models based on data collection, care professionals' input, and iterative testing with older users. These frameworks were selected because their modular architecture provides direct control over speech recognition, dialogue management, and speech synthesis. This separation enables the proposed interaction method to be implemented under consistent, reproducible conditions, with deterministic task handling and clearly attributable system behavior. They therefore provide a suitable experimental platform for evaluating the study's central contribution: the design of an open and modular voice-service architecture. All the source code that implements this service has been made freely available to the community via the public repository [28].

V. CASE STUDY

This section illustrates the application of the proposed cloud-based voice service architecture through a companionship service for older adults, developed in collaboration with the health and social care services of two local councils. These services support an aging population with needs ranging from medication reminders to everyday companionship; however, limited budgets and staffing mean that current in-person support does not exceed one hour per user per day. Initial attempts to use commercial virtual assistants exposed important limitations. Their speech and dialogue models were not sufficiently adapted to older users, while vendor policies restricted changes to elements such as voice configuration, conversational behavior, and wake words. These constraints motivated a second project iteration based on open hardware and software, with customized voice and dialogue models tailored to the target population. The resulting system comprises the SAM Device, which provides the user-facing intelligent-assistant functionality, and the SAM Voice Service, which provides back-end voice and natural-language-processing capabilities. The following subsections describe the device design, the high-availability deployment architecture, and the development environment supporting the service. A dedicated quantitative evaluation of the system, including performance metrics, usability indicators, and statistical treatment of user-acceptance data, is being developed as a separate study. The present case study is not a complete experimental validation but rather an architectural feasibility study.

1. Hardware Prototype of the IoT Device (SAM Device)

The Assistant front-end interface is the VUI provided by an IoT device, the SAM Device, through which older adults interact with the system. This device was designed to accompany older adults, although it may also be highly useful for family members who wish to access the companionship service's features via the voice interface. Figure 4 shows a conceptual block diagram of the device architecture.

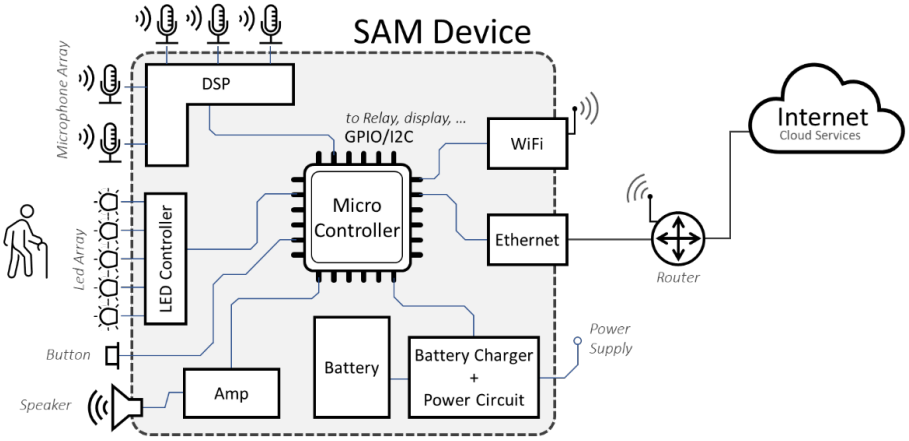


Figure 4. SAM Device high-level hardware architecture.

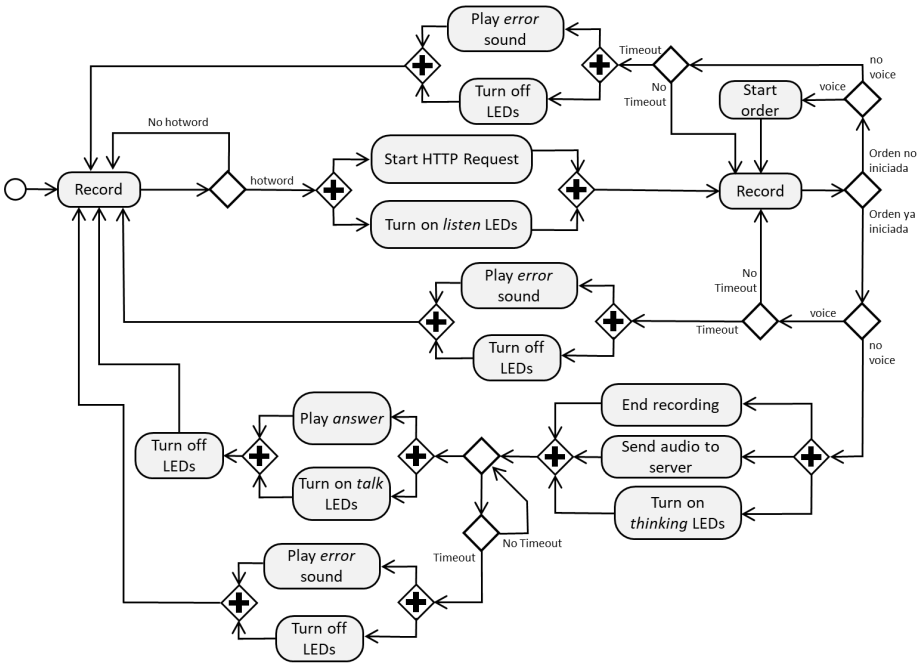


Figure 5. Flowchart of interaction with the SAM Device.

Figure 5 summarises the interaction flow of the SAM Device. The device remains in a passive local-listening state until the wake word is detected, then captures the user's utterance, sends the audio request to SAM Voice Service, plays the returned response, and returns to passive listening. Three different hardware prototypes were built: one based on the ReSpeaker Core v2 board; another based on a Raspberry Pi 3 together with a ReSpeaker 4-Mic Array board; and a final prototype based on a Raspberry Pi and a ReSpeaker 2-Mic Array board. All the prototypes incorporate the hardware components defined in the conceptual architecture (Figure 3).

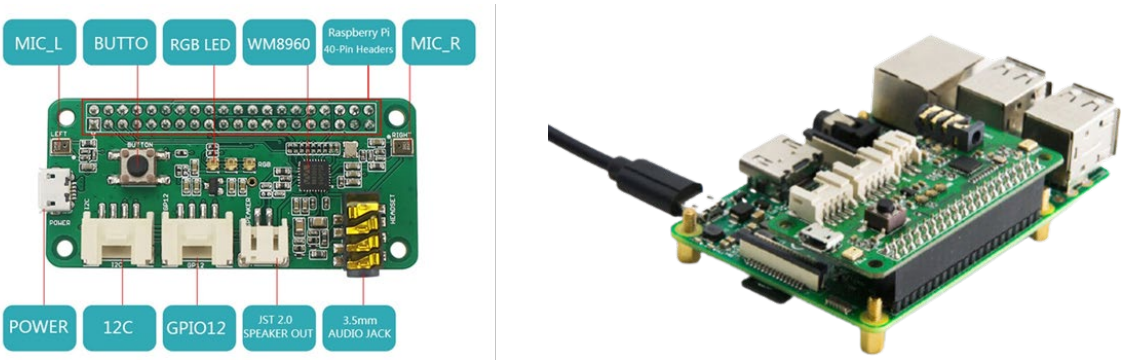


Figure 6. Prototype based on Raspberry Pi 3 Model B+ together with two microphones, four LEDs, speaker, temperature and humidity sensor, and relay.

Figure 6 shows, by way of example, one of the prototypes assembled from a Raspberry Pi 3 Model B+ and an array of microphones and LEDs. All the components incorporated and all the design work carried out followed open-hardware principles. In this way, the proposed hardware device can be easily replicated, maintained, and modified, even if the specific components described here are unavailable.

All the source code for the device is available to the whole community through a public repository [29].

2. Cloud Voice Service Platform Deployment Architecture and Performance Scenario

The service must support both small deployments and larger installations with variable demand. Figure 7 presents the high-availability deployment architecture, in which containerized service instances are deployed in Kubernetes clusters, with Nginx serving as a reverse proxy and load balancer. This design allows individual microservices and storage resources to be scaled independently according to demand and observed performance requirements. Considering the proposed high-availability architecture, a development scenario was constructed as an example (see Figure 8) to achieve the high availability and ease of scaling required in a production environment. This high-availability scenario is largely based on server clusters, storage networks, and load balancers.

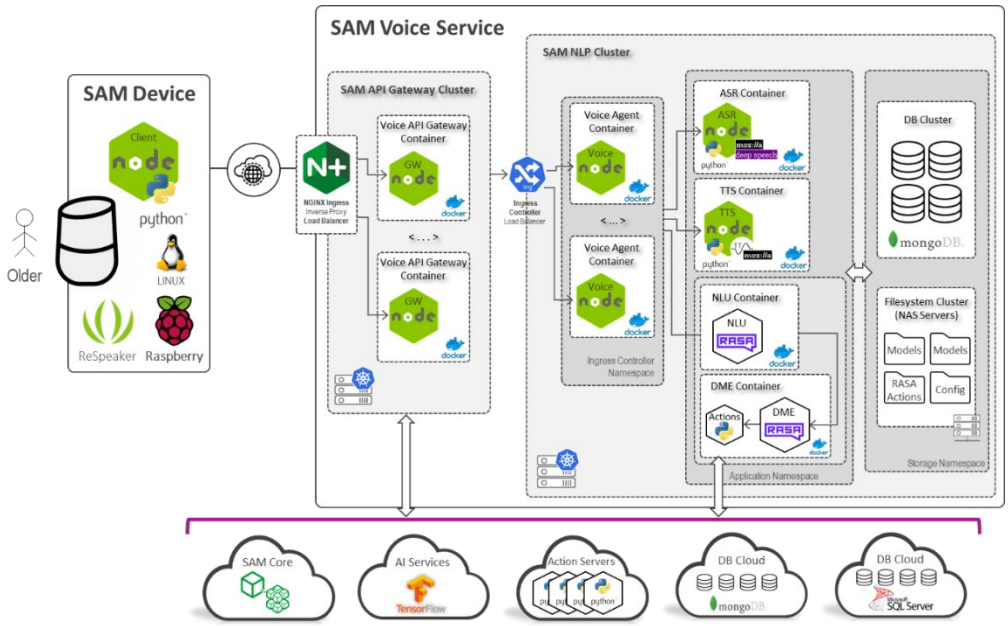


Figure 7. High-availability deployment architecture and the technologies involved.

The feasibility case involved 250 older adults who completed a post-use questionnaire after a mean trial exposure of 21.4 days; 89.2% used SAM daily or several times per week. The sample included participants aged 65–74, 75–84, and 85+, with 51.2% living alone, 79.2% reporting low or basic digital literacy, and 65.6% having no previous voice-assistant experience. Participants also represented different levels of care support and self-reported sensory, mobility, or mild cognitive limitations.

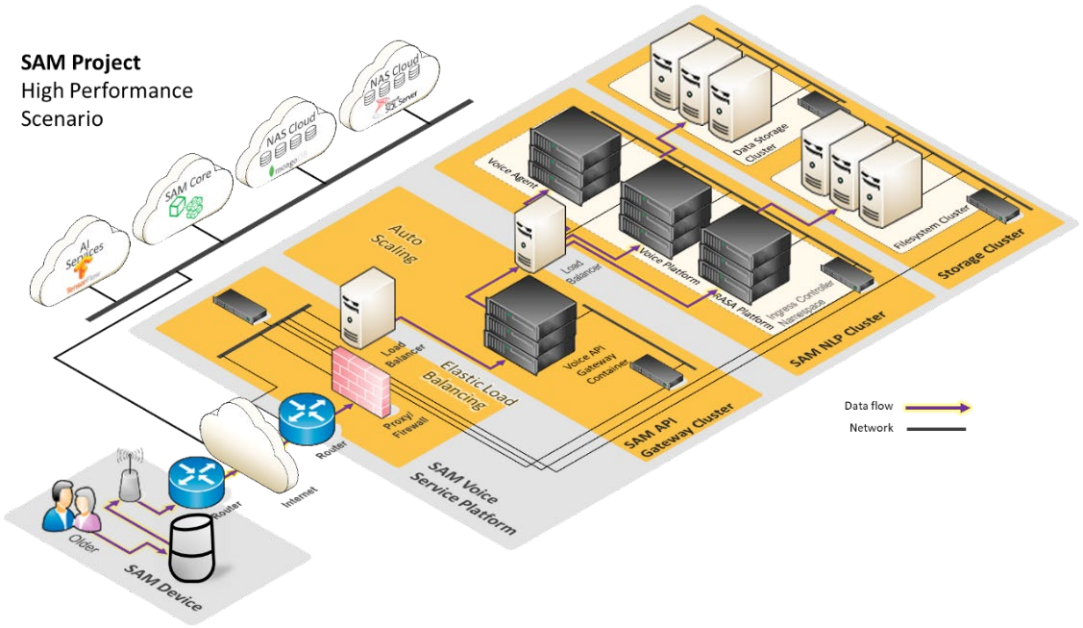


Figure 8. High availability performance scenario.

The questionnaire covered usefulness, ease of use, voice interaction quality, trust and safety, companionship value, and overall acceptance. Positive acceptance was observed in 217 participants (86.8%), with an approximate 95% confidence interval of 82.6–91.0%. The complete aggregated survey results are available as supplementary material in the public repository referenced in [28].

VI. CONCLUSIONS

This paper presented a modular, cloud-based voice-service architecture for Voice User Interfaces tailored to older adults. The proposed architecture spans conceptual, technical, and deployment levels, enabling the integration of open-source components and technologies from different providers. It is intended to support deployments that can be adapted to local functional requirements, available resources, and operational constraints. The architecture was instantiated in a case study on companionship services involving two local-authority social-care initiatives. The resulting system combines SAM Devices, which provide the user-facing interaction, with SAM Voice Service, which supplies back-end voice and natural-language-processing capabilities. The case study illustrates how these components can be deployed and scaled independently to support different service scenarios. The proposed system is intended to complement, rather than replace, family members and health and social care professionals by providing an accessible, voice-based channel for support and interaction. The source code and associated materials have been made available through the public repositories referenced in this paper. Ongoing complementary work will quantitatively evaluate usability, reliability, performance, and service outcomes in broader real-world deployments. This work has limitations. Its performance depends on the availability and population representation of ASR training data for the target population, and adaptation may be needed for different accents, speech characteristics, or deployment settings. Although the architecture supports scalable deployment, cloud processing and high-availability infrastructure entail computational, energy, and maintenance costs. Finally, the case study represents an architectural feasibility case conducted in a specific local context. A dedicated quantitative evaluation is being developed separately, and future deployments should also be validated with their own target users and operational conditions.

REFERENCES

[1]. Eurostat (2026). Demography of Europe: 2026 edition. European Commission. Available at: https://ec.europa.eu/eurostat/web/interactive-publications/demography-2026 (Accessed: 12 June 2026).
[2]. Eurostat (2026). Population projections in the EU. EUROPOP2025 population projections; data extracted April 2026. European Commission. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_projections_in_the_EU (Accessed: 12 June 2026).
[3]. AAL Association (2025). AAL Legacy Study: Final Report. Available at: https://www.aal-europe.eu/wp-content/uploads/2025/10/AAL-Legacy-Final-Report-AAL-consolidated.pdf (Accessed: 13 June 2026).
[4]. OECD and European Commission (2024). Health at a Glance: Europe 2024: State of Health in the EU Cycle. OECD Publishing, Paris. https://doi.org/10.1787/b3704e14-en.
[5]. Lindberg B, Nilsson C, Zotterman D, Söderberg S, Skär L (2013). Using Information and Communication Technology in Home Care for Communication between Patients, Family Members, and Healthcare Professionals: A Systematic Review, Int J Telemed Appl, 2013, no.461829.
[6]. Wagner S, Hunnerup E (2018). Ambient assisted living ecosystem for supporting senior citizens' human system interaction, Proceedings - 2018 11th International Conference on Human System Interaction, HIS, 2018, no. 8431357, 221-225.
[7]. Demiris G (2006). The diffusion of virtual communities in health care: concepts and challenges, Patient Education and Counseling, 2006, 62(2), 178–188.
[8]. Prognos AG (2022). Final Evaluation of the Active and Assisted Living Research and Development Program (AAL2): Final Report. Prepared for the European Commission. Available at: https://www.aal-europe.eu/wp-content/uploads/2024/02/Annex-3a-CONFIDENTIAL-AAL2Evaluation_final.pdf (Accessed: 20 June 2026).
[9]. Son G, Kwon S, LIM Y (2017). Speech Rate Control to Improve Elderly Speech Recognition on Smart Devices. Advances in Electrical and Computer Engineering, 2017, 17, 79-84.
[10]. Pedrozo Campos Antunes T, Souza Bulle de Oliveira A, Hudec R, Brusque Crocetta T, et al (2019). Assistive technology for communication of older adults: a systematic review, Aging and Mental Health, 2019, 23 (4), 417-427.
[11]. Balakrishnan, S., Salim, S. S. B., and Hong, J. L. (2012). User Centered Design Approach for Elderly People in Using Website, Proceedings of the 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT 2012), pp. 382-387.
[12]. Granata C, Pino M, Legouverneur G, Vidal JS, Bidaud P & Rigaud AS (2013). Robot Services for the Elderly with Cognitive Impairment: Testing Usability of Graphical User Interfaces. Technology & Health Care (21:3), pp 217-231.
[13]. Zajicek M (2004). Successful and Available: Interface Design Exemplars for Older Users, Interacting with Computers (16:3), pp. 411-430.
[14]. Klimova B, Poulova P (2018). Older People and Technology Acceptance. In: Zhou J., Salvendy G. (eds) Human Aspects of IT for the Aged Population. Acceptance, Communication, and Participation. ITAP 2018. Lecture Notes in Computer Science, 2018, 10926, 85-94.
[15]. Heerink M, Krose B, Evers V & Wielinga B (2006). The Influence of a Robot's Social Abilities on Acceptance by Elderly Users. ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006, pp. 521-526, doi: 10.1109/ROMAN.2006.314442.
[16]. Esposito A, Amorese T, Cuciniello M, Riviello MT, Esposito AM, Troncone A, and Cordasco G (2019). The Dependability of Voice on Elders' Acceptance of Humanoid Agents. INTERSPEECH 2019 September 15–19, 2019, Graz, Austria. https://www.isca-speech.org/archive/Interspeech_2019/pdfs/1734.pdf.
[17]. The Top 82 Text To Speech Open Source Projects, https://awesomeopensource.com/projects/text-to-speech, 2020.

[18]. Maciá Fiteni, Àlex (2021). SAM Device. Dispositivo IoT para la asistencia a mayores. Grado en Ingeniería Multimedia - Trabajos de Fin de Grado. Repositorio Institucional de la Universidad de Alicante. 25 de septiembre de 2021. http://hdl.handle.net/10045/118144.
[19]. MacCartney B. (2014). Understanding Natural Language Understanding, ACM SIGAI Bay Area Chapter Inaugural Meeting.
[20]. Grindrod KA, Li M & Gates A (2014). Evaluating User Perceptions of Mobile Medication Management Applications with Older Adults: A Usability Study. JMIR Mhealth Uhealth, (2:1), e11.
[21]. Sutter C, and Müsseler J (2007). User Specific Design of Interfaces and Interaction Techniques: What Do Older Computer Users Need? UAHCI 2007 Proceedings, pp 1020-1029, Beijing, China.
[22]. Gregor P, Newell AF & Zajicek M (2002). Designing for Dynamic Diversity: Interfaces for Older People, Proceedings of the 5th International ACM Conference on Assistive Technologies, pp 151-156, Edinburgh, Scotland.
[23]. Glisky EL (2007). Changes in Cognitive Function in Human Aging, In Brain Aging: Models, Methods, and Mechanisms, pp 4-20, CRC Press.
[24]. Dodd C, Athauda R & Adam MTP (2017). Designing User Interfaces for the Elderly: A Systematic Literature Review. Conference: Australasian Conference on Information Systems. Hobart, Australia. 2017.
[25]. Queirós, A., Silva, A., Alvarelhão, J., Rocha, N. P. & Teixeira, A. (2015). "Usability, Accessibility and Ambient-assisted Living: A Systematic Literature Review," Universal Access in the Information Society (14:1), pp 57-66.
[26]. Erl T (2009) SOA Design Patterns. Prentice Hall PTR; 1st Eds, New York, January 2009. ISBN 978-0136135166.
[27]. Richardson C (2018). Microservices Patterns. Manning. October 2018. ISBN 978-1617294549.
[28]. Àlex Maciá-Fiteni (2022). GitHub Repository. https://github.com/AlexMaciaFiteni/OpenSourceVoiceService.
[29]. Àlex Maciá-Fiteni (2022). GitHub Repository. https://github.com/AlexMaciaFiteni/OpenSourceHotwordDetection.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Evaluating the Effectiveness of the AIM-5 Model for AI Teaching Self-Efficacy and AI Integration Readiness

Medina Bokan
Department of Informatics and Digitalization of Education
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
bokanmadina98@gmail.com
ORCID: 0000-0003-0767-4094

M. Tokanov
Department of Information Technology and Artificial Intelligence
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
mansur_tokanov@mail.ru
ORCID: 0000-0003-3679-5485

Zh. Zhiyembayev
Department of Informatics and Digitalization of Education
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
jomart73@mail.ru
ORCID: 0009-0005-0438-4077

D. Shalkarkyzy
Department of Information Technology and Artificial Intelligence
Zhetysu University named after I. Zhansugurov
Taldykorgan, Kazakhstan
shalkarkyzyd@gmail.com
ORCID: 0009-0007-9480-0621

Abstract
AI is increasingly embedded in Kazakhstan's education system, creating demand for future informatics teachers who are both AI-literate and confident in applying AI in practice. This quasi-experimental study (pre-test/post-test, non-equivalent control group) evaluated the effect of the AI Integration Methodology 5-stages (AIM-5) model on AI Teaching Self-Efficacy and AI Integration Readiness among 84 third-year informatics students (EG, n=44; CG, n=40) at Abai Kazakh National Pedagogical University during the 2025–2026 academic year. The EG completed a 15-week AIM-5 program embedded in a methodology course; the CG followed the standard curriculum. Both outcomes were measured with validated, translated scales (α=0.85; α=0.83). Groups were statistically equivalent at pre-test (p=0.90). At post-test, the EG showed significantly greater gains than the CG in AI Teaching Self-Efficacy (12.2% vs 2.7%; t(82)=2.50, p=0.015, d=0.55) and AI Integration Readiness (10.4% vs 2.0%; t(82)=2.52, p=0.014, d=0.55). These small-to-moderate but reliable gains suggest AIM-5 supports early psychological readiness for AI integration, though longer, field-embedded interventions may be needed to produce deeper competency change.

Keywords: AI teaching self-efficacy; AI integration readiness; AIM-5 model; future informatics teachers; digital literacy; quasi-experimental study; TPACK-AI; teacher education.

I. INTRODUCTION

Generative AI's rapid adoption in global education has reshaped the competencies expected of teachers. For future informatics teachers, AI now occupies a "dual reality": it is both a content area to be taught and a methodological tool to be applied in teaching and learning practice. Bridging these two roles requires more than declarative knowledge of AI. It requires the psychological readiness to act, namely, teaching self-efficacy and integration readiness. Bandura's social cognitive theory identifies self-efficacy as a key factor mediating behavior, such that teachers with stronger self-efficacy beliefs are more persistent, more willing to try new teaching methods, and more resilient when they encounter technical failures [1]. The most widely used measure of this construct is the Teachers' Sense of Efficacy Scale (TSES) by Tschannen-Moran and Woolfolk Hoy [2], which has been validated to predict classroom innovation and the adoption of new instructional technologies [3].

To address teachers' needs in the AI era, Chiu et al. [4] introduced the Teacher AI Competence Self-Efficacy (TAICS) construct, which reflects teachers' confidence in employing, integrating, and controlling AI tools in teaching. Additionally, Çelik Demirci et al. [5] validated the Brief General AI Self-Efficacy Scale with 969 pre-service teachers in Turkish universities, further supporting the importance of AI self-efficacy in teacher preparation. The construct of AI integration readiness is complementary but distinct. Wang et al. [6] proposed four interrelated dimensions of teacher AI readiness — cognition, ability, vision, and ethics — and showed that all four were positively related to AI tool use in a sample of 3,164 primary school teachers. Aksoy et al. [7] subsequently developed the Readiness for AI Applications Scale (RAIS), comprising technology self-efficacy, interaction with students, and ethical awareness subscales, which was later validated in Turkish by Özüdoğru and Yildiz Durak [8]. These instruments collectively operationalize the UNESCO AI competency framework for teachers [9], which is grounded in the human-centered, ethical, and pedagogically informed use of AI in the classroom. Contextual evidence also comes from Turkish research on the psychological challenges of AI implementation. Ayduğ and colleagues [10] found that Turkish pre-service teachers report high digital literacy but moderate AI anxiety, with the technical sub-dimension of digital literacy accounting for 10.3% of variance in AI anxiety. Akdeniz and Özdınç [11] surveyed AI-in-education research in Türkiye and identified the same fragmentation observed elsewhere: emerging tools without systematic pedagogical models for teacher preparation. These findings parallel the situation in Kazakhstan, where AI tools are increasingly available but structured curricular integration remains limited [12]. Kazakhstan's policy landscape has decisively turned toward AI in teacher training. The Presidential Address "Kazakhstan in the Age of Artificial Intelligence" [13] called for AI literacy to be integrated into school curricula and for pedagogical universities to align graduate competencies with an AI-integrated labor market. Within this policy, the primary-school subject "Digital Literacy" holds strategic importance, and the informatics teachers who will teach it must be both technically competent and pedagogically prepared to apply AI tools appropriately. Yet the systematic review by Zawacki-Richter et al. [14] found that only a quarter of teacher education programs worldwide offer pedagogical modules specifically on AI. In our earlier quasi-experimental study [15], future informatics teachers showed substantial gains in AI literacy and motivational engagement with the AIM-5 model. However, such gains do not necessarily translate into confidence and readiness to implement AI in real classrooms: a student may know what AI is and still feel unprepared to use it in teaching [16]. The present study extends that earlier work by testing whether the same AIM-5 intervention, delivered as a full 15-week program within the course "Methodology of Teaching the Subject 'Digital Literacy' in Primary School," also improves two psychological pre-conditions of AI integration: AI Teaching Self-Efficacy and AI Integration Readiness. AIM-5 has three distinguishing features from generic TPACK-AI training and stand-alone AI-literacy modules. First, AIM-5 doesn't involve one-off training sessions; it's embedded in a full-semester methodology course and delivered in a sequence of five stages, from orientation to independent classroom application, providing AI literacy as declarative knowledge. Second, the psychological antecedents of AI use (self-efficacy and readiness) are targeted rather than just technical AI competence, which Wang and Chuang [16] have identified as a gap between knowledge and preparedness to use AI. Third, each step of AIM-5 results in a tangible, evaluated pedagogical artifact: a tool-comparison worksheet, an AI-supported activity portfolio, a TPACK-AI plan, a microteaching recording, and a reflection journal, providing the model with an assessment framework that is not always found in general AI-literacy interventions. The weekly activities, assessment, and AI tools used for each stage are described.

The research questions are:

1. Does participation in a 15-week AIM-5 program produce greater gains in AI Teaching Self-Efficacy among future informatics teachers than the standard curriculum?

2. Does the AIM-5 model produce greater gains in AI Integration Readiness? We hypothesized that the EG would show modest but reliable gains on both outcomes, consistent with Bandura's prediction that belief-based constructs change slowly even under structured intervention.

II. MATERIALS AND METHODS

1. Research design

The quasi-experimental pre-test/post-test non-equivalent control group design was used [17]. Random assignment of individual students was not feasible because the intervention was embedded in an existing methodology course delivered to two intact third-year class groups of the “Informatics” program; students were already enrolled in these groups for administrative and timetabling reasons that predate the study. One intact group was designated the experimental group (EG, n=44) and the other the control group (CG, n=40). Because assignment followed pre-existing class rosters rather than random allocation, causal claims are treated cautiously throughout this paper, and pre-test equivalence between groups was formally verified (Section III) before any between-group comparison was interpreted. The research took place at Abai Kazakh National Pedagogical University in Almaty, Kazakhstan, in the 2025–2026 academic year.

2. Participants and pre-experiment baseline survey

The total sample comprised 84 third-year future informatics teachers. All 84 students completed an AI self-efficacy pretest before the intervention. The questionnaire was administered online in Kazakh and Russian and comprised 18 self-efficacy items rated on a 5-point Likert scale (1–5). The two items most relevant to AI integration in informatics teaching are reported in Figure 1: (1) students’ self-reported competence in selecting suitable AI tools for informatics lessons, and (2) students’ self-reported competence in creating informatics learning materials, tasks, and digital resources using AI. Both items indicated moderate, polarised baseline self-efficacy, supporting the case for a structured intervention.

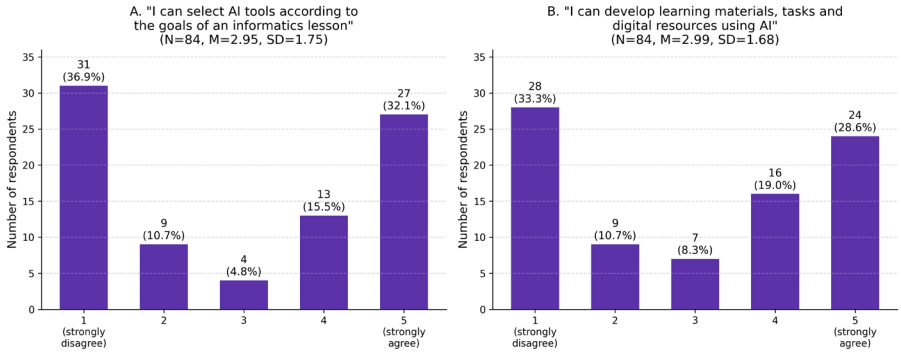


Figure 1. Pre-experiment baseline self-efficacy: distribution of responses on the two key items administered to all 84 participants before the intervention (5-point Likert scale).

3. Ethical considerations

Ethical considerations for the research procedures in this study were addressed in accordance with the ethical guidelines of Abai Kazakh National Pedagogical University, and permission to conduct the study was granted before data collection. All participants were given an informed consent form, which explained the purpose of

the study, the voluntary nature of participation, and that they could withdraw from the study at any time without penalty or impact on their academic standing, as well as the confidentiality and anonymization of their responses, before completion of the pre-test survey. None of the data was collected directly identifying, and all the data were analyzed and reported at the group level only.

4. Intervention: the 15-week AIM-5 program

The experimental group completed a 15-week AIM-5 course embedded in "Methodology of Teaching the Subject 'Digital Literacy' in the Primary School." AIM-5 is a five-stage pedagogical design model extending Technological Pedagogical Content Knowledge (TPACK) [18] to include AI. The weekly focus, activities, and assessed tasks for each stage are described below.

1. AI Orientation (Weeks 1–3). Weekly seminars introduced generative AI tools (ChatGPT, Claude, Microsoft Copilot), their capabilities and limitations, and age-appropriate use in primary education; students completed structured tool-comparison worksheets. Assessed task: a one-page "AI capability map" comparing three tools against age-appropriateness criteria.

2. AI Task Engagement (Weeks 4–6). Students used AI tools to generate and then critically revise draft digital-literacy activities (quizzes, worksheets, short explanatory texts) for 7–10-year-olds, evaluating outputs for accuracy, age-appropriateness, and pedagogical alignment. Assessed task: a portfolio of two AI-assisted activities with a written critique of the AI-generated content's limitations.

3. Principles of AI Methodological Design (Weeks 7–9). Working in pairs, students created full lesson plans for "Digital Literacy" using a TPACK-AI planning matrix that maps AI tool use onto content, pedagogy, and technology dimensions. Assessed task: a lesson plan scored against a TPACK-AI rubric by the course instructor.

4. Practicum Simulation (Weeks 10–12). Students used AI (primarily Claude and Microsoft Copilot) as a co-design partner to prepare 15-minute microteaching episodes, which were delivered to peers in a simulated primary classroom setting, followed by structured peer feedback. Assessed task: a video-recorded microteaching episode with peer- and instructor-rated rubric scores.

5. Reflective Synthesis (Weeks 13–15). Students kept a structured weekly reflection journal on what AI contributed to their planning, where it fell short, and the ethical considerations — data privacy, bias, and over-reliance — raised by using AI with children, and closed with a group reflection seminar—assessed task: a final reflective journal, graded for depth of critical engagement rather than length.

All AI tools used in the program (ChatGPT, Claude, and Microsoft Copilot) were pedagogically filtered, and their outputs had to be critically evaluated, adapted, and contextualized by students before use in a lesson. The control group followed the same base course but did not engage in the systematic, scaffolded use of AI described above.

5. Intervention fidelity

To support consistent delivery, all EG sessions were taught by the same instructor using a shared, lesson-by-lesson facilitation guide developed specifically for the AIM-5 program, and instructional materials (slides, worksheets, and rubrics) were standardized across the 15 weeks. Attendance was recorded at every session, and a second researcher independently observed a sample of sessions across the five stages using a structured fidelity

checklist to confirm that planned activities, assessed tasks, and AI tools matched the intended protocol; no material deviations from the planned curriculum were identified during these observations.

6. Instruments

AI Teaching Self-Efficacy was measured with a 15-item scale adapted from the Teachers' Sense of Efficacy Scale (TSES) [2] and the Teacher AI Competence Self-Efficacy (TAICS) scale [4]. Items were rated on a 5-point Likert scale. They addressed three sub-domains: efficacy of instructional strategies, classroom management, and student engagement, each contextualized to AI use in primary-school digital literacy lessons. Reliability was satisfactory in the pilot sample (Cronbach's $\alpha = 0.85$). AI Integration Readiness was measured using a 16-item scale adapted from the Readiness for AI Applications Scale (RAIS) [7] and its Turkish adaptation by Özüdoğru and Yildiz Durak [8], covering technology self-efficacy, interaction with students, and ethical awareness (Cronbach's $\alpha = 0.83$). Both instruments were translated into Kazakh and Russian, back-translated for equivalence, and content-validated by three experts in informatics pedagogy and AI in education (Content Validity Index, CVI = 0.88).

7. Data analysis

Descriptive statistics (means, standard deviations) were computed for both groups at pre- and post-test. Between-group differences at pre- and post-test were tested with independent-samples t-tests, and Cohen's d was used to estimate effect sizes. Within-group change was tested with paired-samples t-tests. The significance level was set at $\alpha = 0.05$. Because the sample size is modest and several effects are small to medium, statistical significance was interpreted alongside effect size rather than in isolation, to avoid overstating the practical importance of small mean differences. Statistical analyses were conducted in SPSS v.27.

III. RESULTS

Figure 1 shows the distribution of responses to two key baseline self-efficacy items administered to all 84 participants before the intervention. Average scores of 2.95 and 2.98 points (1–5 scale) with bimodal response patterns indicate that the group was initially heterogeneous and polarised: about a third of students rated themselves at the lowest self-efficacy level, while about 30% rated themselves at the highest. This pattern is consistent with the early-implementation stage of AI in informatics teacher training described for both Kazakhstan [12] and Turkey [11].

At pre-test, the experimental and control groups were statistically equivalent on AI Teaching Self-Efficacy (EG: M=2.96, SD=0.74; CG: M=2.94, SD=0.72; t(82)=0.13, p=0.90) and on AI Integration Readiness (EG: M=2.99, SD=0.76; CG: M=2.97, SD=0.73; t(82)=0.12, p=0.90), confirming group equivalence at baseline despite non-random assignment.

Table 1. AI Teaching Self-Efficacy: descriptive statistics and within-group comparisons

Group	n	Pre-test M (SD)	Post-test M (SD)	Δ (abs)	Δ (%)	Paired t (df)	p	d
EG	44	2.96 (0.74)	3.32 (0.50)	+0.36	+12.2%	3.65 (43)	<0.001	0.55
CG	40	2.94 (0.72)	3.02 (0.60)	+0.08	+2.7%	0.76 (39)	0.45	0.12

Note. Scale range: 1–5. Between-group post-test comparison: t(82)=2.50, p=0.015, d=0.55. Pre-test equivalence between groups: t(82)=0.13, p=0.90.

Figure 2 presents AI Teaching Self-Efficacy scores at pre- and post-test by group. The EG showed a within-group increase of 0.36 points (from M=2.96 to M=3.32, a 12.2% relative gain), which was statistically significant (paired t(43)=3.65, p<0.001, d=0.55). The CG showed a much smaller, non-significant increase of 0.08 points (from M=2.94 to M=3.02, a 2.7% relative gain; paired t(39)=0.76, p=0.45). The between-group post-test difference was statistically significant with a medium effect size: t(82)=2.50, p=0.015, Cohen's d=0.55. Although statistically reliable, this effect size and the modest absolute gain (0.36 points on a 5-point scale) indicate an early, incremental shift in self-efficacy rather than a large practical change in teaching confidence; detailed statistics are presented in Table 1.

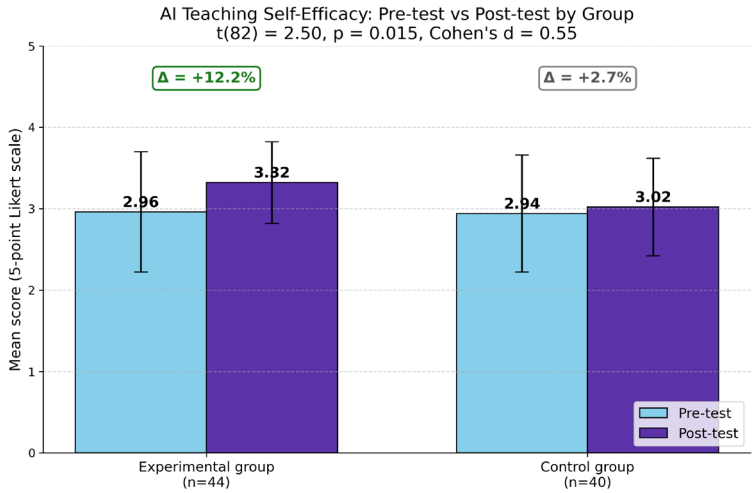


Figure 2. AI Teaching Self-Efficacy: pre-test and post-test by group (EG, n=44; CG, n=40). Error bars represent ±1 SD. Scale range: 1–5.

Table 2. AI Integration Readiness: descriptive statistics and within-group comparisons

Group	n	Pre-test M (SD)	Post-test M (SD)	Δ (abs)	Δ (%)	Paired t (df)	p	d
EG	44	2.99 (0.76)	3.30 (0.45)	+0.31	+10.4%	3.11 (43)	0.003	0.47
CG	40	2.97 (0.73)	3.03 (0.53)	+0.06	+2.0%	0.58 (39)	0.57 ns	-

Note. Scale range: 1–5. Between-group post-test comparison: t(82)=2.52, p=0.014, d=0.55. Pre-test equivalence between groups: t(82)=0.12, p=0.90.

Figure 3 presents the AI Integration Readiness results, which followed a parallel pattern. The EG showed a 0.31-point gain (from M=2.99 to M=3.30, a 10.4% relative gain; paired t(43)=3.11, p=0.003, d=0.47), while the CG showed only a 0.06-point gain (from M=2.97 to M=3.03, a 2.0% relative gain; paired t(39)=0.58, p=0.57). The between-group post-test difference was statistically significant: t(82) = 2.52, p = 0.014, Cohen's d = 0.55. As with AI Teaching Self-Efficacy, this medium effect size should be read as evidence of a real but modest difference rather than a large practical change; detailed results are summarised in Table 2.

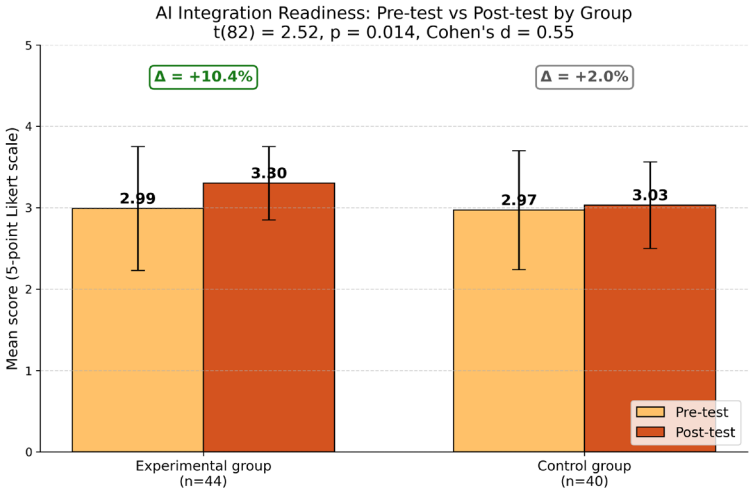


Figure 3. AI Integration Readiness: pre-test and post-test by group (EG, n=44; CG, n=40). Error bars represent ±1 SD. Scale range: 1–5.

IV. DISCUSSION

Two patterns merit interpretation. First, the magnitude of the EG gains (around 10–12%) is smaller than the gains reported in our previous study on AI literacy and motivation [15], in which the same AIM-5 model produced larger effects ($d > 1.0$). This distinction may be theoretical: AI literacy and motivation are knowledge- and interest-based constructs that change relatively quickly under structured exposure, whereas self-efficacy and readiness are belief-based constructs that change more slowly and depend on mastery experiences. A 15-week intervention likely provides only a few authentic mastery experiences in a real primary school setting, thereby limiting potential gains. While statistically significant, the 12.2% increase in AI Teaching Self-Efficacy should be read as a preliminary shift rather than a change in teacher identity — a pattern that echoes Çelik Demirci et al. [5], who found AI knowledge to be the strongest correlate of AI self-efficacy among Turkish pre-service teachers. It is also important not to over-interpret these effects. Although both between-group comparisons reached statistical significance at $\alpha=0.05$, the corresponding effect sizes ($d=0.55$ for both outcomes) are medium by conventional benchmarks, and the absolute score gains (0.36 and 0.31 points on a 5-point scale) are modest in real-world terms. In a sample of this size, statistical significance indicates that a real, replicable difference of moderate size was detected. It does not, by itself, demonstrate a large or transformative practical effect on teachers' classroom-ready AI competence. Second, for both outcomes, there was a slightly larger gain in self-efficacy than in readiness, consistent with the conceptual model proposed by Wang et al. [6] and empirically supported by the Turkish adaptation of Özüdoğru and Yildiz Durak [8], in which readiness incorporates self-efficacy plus ethical and visionary components that take longer to develop. The 15-week AIM-5 program likely under-samples the full impact of the Reflective Synthesis stage, which engages students in ethical discussions about AI use in primary classrooms; this stage is the most plausible mechanism by which the intervention influences readiness. A similarly small, non-significant improvement was observed in the control group on both measures. This is likely explained in part by the wider institutional context: students already use AI tools on their own devices, and the university offers optional AI-related workshops [12]. AI appears to be diffusing into students' informal learning even outside deliberately AI-focused courses, producing a small "natural drift" effect. This parallels Ayduğ et al.'s [10] finding that Turkish pre-service teachers' attitudes shift modestly with mere exposure to AI, but that substantial, structured pedagogical exposure is needed to produce educationally meaningful change in efficacy and preparedness. Viewed through the UNESCO AI Competency Framework for

Teachers [9], the present findings suggest that AIM-5 is a promising but limited intervention: it appears to foster the technological and pedagogical components of AI competence at the acquisition level, while the deepening and creation levels would likely require longer, field-embedded extensions of the model with sustained ethical leadership.

The findings here are limited in their interpretations by several factors, which we explicitly state and do not make a formality of. First, the sample is quite small (N=84) and drawn from a single cohort at a single university. Groups were not randomly assigned but rather followed existing class schedules, so the results should be viewed as preliminary and hypothesis-generating rather than definitive proof of a causal link between AIM-5 and the gains observed, and are not generalizable beyond similar institutional settings without replication. Second, the 15-week duration is still a full semester, but it is short relative to the period during which self-efficacy beliefs are known to remain stable; thus, gains may not be sustained. Third, the study is based on self-reported self-efficacy and readiness; behaviors would need to be observed, such as observed AI interaction during teaching practice, to support causal and practical claims. Finally, non-random assignment of students to groups poses a risk of selection bias, which is mitigated by the groups' equivalence in the pre-study assessment. Still, unmeasured differences between the intact class groups (e.g., instructor effects before the study, informal peer influence) are not eliminated.

V. CONCLUSION

We tested the effectiveness of the AI Integration Methodology 5-stages (AIM-5) model in improving AI Integration Readiness and AI Teaching Self-Efficacy among future informatics teachers, based on a quasi-experimental study involving 84 third-year students in the course “Methodology of Teaching the Subject ‘Digital Literacy’ in Primary School” at Abai Kazakh National Pedagogical University. The experimental group showed statistically significant, small-to-medium gains relative to the control group in AI Teaching Self-Efficacy (12.2% vs. 2.7%; $t(82)=2.50$, $p=0.015$, $d=0.55$) and AI Integration Readiness (10.4% vs. 2.0%; $t(82)=2.52$, $p=0.014$, $d=0.55$); given the study's sample size and non-random design, these results should be read as preliminary evidence rather than definitive proof of effectiveness. Three implications follow. First, a 15-week AIM-5 intervention appears sufficient to induce small-to-moderate changes in the psychological pre-conditions for AI integration, but longer, field-embedded extensions may be necessary to produce transformative change in teacher identity. Second, the parallel increases in self-efficacy and readiness support the conceptual accounts of Wang et al. [6] and Chiu et al. [4], and align with recent Turkish studies [5], [8], [10], indicating that these constructs develop on different timelines and require different pedagogical supports. Third, the small but persistent improvements in the control group underline the need to benchmark structured AI-integration models against realistic, AI-exposed baselines rather than AI-free comparison groups. Future research could pursue three directions: (1) longitudinal tracking of AIM-5 participants during their first two years of teaching practice; (2) multi-site replication using the same instruments across pedagogical universities in Central Asia, Türkiye, and other regions, with randomized assignment where institutionally feasible; and (3) incorporation of behavioral measures to triangulate the self-report results presented here.

ACKNOWLEDGEMENT

The authors thank the third-year students of the educational program “Informatics” at Abai Kazakh National Pedagogical University for their participation, and the three subject-matter experts who validated the adapted instruments used in this study.

REFERENCES

[1]. A. Bandura, *Self-Efficacy: The Exercise of Control.* New York, NY, USA: W.H. Freeman, 1997.
[2]. M. Tschannen-Moran and A. Woolfolk Hoy, "Teacher efficacy: Capturing an elusive construct," *Teaching and Teacher Education*, vol. 17, no. 7, pp. 783–805, 2001, doi: 10.1016/S0742-051X(01)00036-1.
[3]. P. A. Ertmer and A. T. Ottenbreit-Leftwich, "Teacher technology change: How knowledge, confidence, beliefs, and culture intersect," *Journal of Research on Technology in Education*, vol. 42, no. 3, pp. 255–284, 2010, doi: 10.1080/15391523.2010.10782551.
[4]. T. K. F. Chiu, G. Falloon, Y. Song, V. W. L. Wong, L. Zhao, and M. Ismailov, "Development and validation of the Teacher AI Competence Self-Efficacy (TAICS) scale," *Education and Information Technologies*, 2024, doi: 10.1007/s10639-024-13094-z.
[5]. S. Çelik Demirci, M. Beşaltı, and Ü. Kul, "Psychometric validation of the Turkish Brief General AI Self-Efficacy Scale: examining its role in AI anxiety, acceptance, and demographic predictors," *Frontiers in Psychology*, vol. 17, art. 1714147, 2026, doi: 10.3389/fpsyg. 2026.1714147. [In press at time of writing; verify final volume/page details before publication.]
[6]. X. Wang, L. Li, S. C. Tan, L. Yang, and J. Lei, "Preparing for AI-enhanced education: Conceptualizing and empirically examining teachers' AI readiness," *Computers in Human Behavior*, vol. 146, art. 107798, 2023, doi: 10.1016/j.chb.2023.107798.
[7]. N. C. Aksoy, E. Kabadayı, Y. Yılmaz, and S. O. Sezer, "AI readiness scale for teachers: Development and validation," *Education and Information Technologies*, 2025, doi: 10.1007/s10639-024-13087-y.
[8]. G. Özüdoğru and H. Yildiz Durak, "Conceptualizing pre-service teachers' artificial intelligence readiness and examining its relationship with AI literacy, digital citizenship, AI-enhanced innovation and perceived threats," *Industrial and Commercial Training*, 2025, doi: 10.1177/02666669251335657.
[9]. F. Miao and M. Cukurova, *AI Competency Framework for Teachers.* Paris, France: UNESCO, 2024. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000391104
[10]. D. Ayduğ et al., "Are Turkish pre-service teachers worried about AI? A study on AI anxiety and digital literacy," *AI & Society*, 2025, doi: 10.1007/s00146-025-02348-0.
[11]. M. Akdeniz and F. Özdınç, "Examination of Turkey-addressed studies regarding artificial intelligence in education," *YYU Journal of Education Faculty*, vol. 18, no. 1, pp. 912–932, 2021, doi: 10.33711/yyuefd. 938734.
[12]. Y. Smagulov, Zh. Zhiyembaev, and M. Bokan, "Artificial intelligence in education: Impact, applications and prospects," *Vestnik KazNPU named after Abai, Pedagogical Sciences Series*, vol. 89, no. 1, pp. 122–135, 2025.
[13]. K.-J. Tokayev, "Kazakhstan in the Age of Artificial Intelligence: Current Issues and Their Resolution through Fundamental Digital Transformation: Address to the People of Kazakhstan," Akorda, Sept. 8, 2025. [Online]. Available: https://www.akorda.kz/kz/
[14]. O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education: Where are the educators?" *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, art. 39, 2019, doi: 10.1186/s41239-019-0171-0.
[15]. M. Ye. Bokan, Ye. Zh. Smagulov, and Zh. T. Zhiembayev, "Effects of AI integration on AI literacy and motivation in future informatics teacher education," *Vestnik of Zhetysu University*, 2025. [In press; verify final citation details before publication.]
[16]. Y. Y. Wang and Y. W. Chuang, "Artificial intelligence self-efficacy: Scale development and validation," *Education and Information Technologies*, vol. 29, no. 4, pp. 4785–4808, 2024, doi: 10.1007/s10639-023-12015-w.
[17]. W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA, USA: Houghton Mifflin, 2002.

[18]. P. Mishra and M. J. Koehler, "Technological pedagogical content knowledge: A framework for teacher knowledge," *Teachers College Record*, vol. 108, no. 6, pp. 1017–1054, 2006, doi: 10.1111/j.1467-9620.2006.00684.x.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Generative AI through the Student Eye: perspectives, attitudes, and usage

Mariam Alshammari
Partner – Education Strategy, Policy, and Leadership
Innotech Consulting,
Kuwait
m.alshammari@innotech.consulting

Fajer Alshammari
English Language Program
Australian University
Mubarak Al-Abdullah Area,
Kuwait
f.alshammari@au.edu.kw

Khuloud Alkhateeb
Computer Science Department
Gulf University for Science and Technology
Mubarak Al-Abdullah Area,
Kuwait
khateeb.k@gust.edu.kw

Fatma Alalawi
Student Research Assistant
Gulf University for Science and Technology
Mubarak Al-Abdullah Area,
Kuwait
GUST0025736@gust.edu.kw

Abstract
The rapid adoption of generative artificial intelligence (AI) tools has transformed higher education, creating new opportunities for learning while raising concerns regarding academic integrity, ethical use, and institutional policy. This study explores university students' perceptions, attitudes, and academic uses of generative AI within a private English-medium university in Kuwait. An online survey consisting of 21 questions was administered to students across all levels of study, yielding 418 responses. The survey examined four key areas: the normalization of AI use in academia; whether students use AI as a support tool or as a substitute for learning; students' understanding of ethical concerns surrounding AI; and their expectations of faculty guidance and institutional policy. Descriptive statistics and Pearson's chi-square analyses were employed to examine relationships between demographic variables and AI use. Results indicate that generative AI has become highly normalized, with 66.7% of students reporting academic use, primarily for research assistance, assignments, writing support, and summarization. While only a small number admitted to directly copying AI-generated content, most students reported attempting to understand AI outputs before incorporating them into their work. Recorded AI use was significantly higher with increasing study level, although no significant differences were observed by gender or academic major. Students demonstrated considerable uncertainty about the ethical boundaries of AI use, yet expressed strong support for clear institutional guidelines and the responsible integration of AI into education. The findings suggest that students view generative AI as a valuable learning aid rather than a replacement for learning itself, highlighting the urgent need for comprehensive AI literacy initiatives, faculty guidance, and institution-wide policies that promote ethical and effective AI use in higher education.

Keywords: Generative Artificial Intelligence; Higher Education; Student Perspectives; Academic Integrity; AI Ethics; Educational Technology; Digital Literacy; ChatGPT; University Policy

I. INTRODUCTION

The application of Artificial Intelligence (AI) in education is a multifaceted topic, and effective understanding requires narrowing in on each stakeholder's perspective. The purpose of this paper is to address the key stakeholders in any educational setting: the student [1]. Therefore, here we try to shed light on students' perceptions, attitudes, and usage patterns of AI in relevance to their academic work. To draw a parallel, there was a time when calculators were protested by educators, who feared that overreliance on such tools would be detrimental to students' arithmetic skills [2], [3]. Today, fear of overreliance on technology is once again rampant, and this time it may feel more valid than ever with the rise of generative AI tools like ChatGPT and its like. Similar to the calculator, the internet, and the many advancements in educational technology we have seen from the mid-1980s to today, mere tools may be overestimated or underestimated in their popularity, use, and educators' ability to control their use [4]. The authors of this paper hold that any tool can be used or misused, and that the onus of guiding its use falls on the educator, not the student. Therefore, we begin our analysis at the proverbial ground zero of the matter: understanding students' uses and the expectations of their institution and educators. Four key research questions are addressed to better frame and guide our understanding as educators:

1. How normalized is the use of generative AI in universities today?
2. Is AI used as a supporting tool, or are students fully dependent on it for answers?
3. Can students of the digital age recognize the ethical concerns associated with AI?
4. Do students expect more guidance on AI from their professors?

II. LITERATURE REVIEW

The use of Generative AI, particularly ChatGPT, has become increasingly embedded within higher education as students employ these tools to complete assignments, essays, reports, projects, and research tasks [5]. The rapid adoption of ChatGPT, which surpassed one million users within its first week of release [6], reflects how quickly generative AI has become integrated into academic environments. Existing research consistently demonstrates that awareness and use of generative AI among university students are widespread, with some studies identifying ChatGPT as the dominant platform for academic purposes [7], [8]. While this growing body of literature establishes that AI use is now commonplace, it also suggests that adoption is driven by a combination of academic, psychological, and technological factors rather than a single motivation. Across studies, efficiency consistently emerges as the primary reason students engage with generative AI. Abbas et al. [5], Chung [7], and Sousa and Cardoso [8] all report that students perceive AI as a means of reducing workload, saving time, improving productivity, and supporting learning through summarization, idea generation, translation, writing assistance, and concept explanation. Although these studies differ in emphasis, they converge on the conclusion that students primarily value generative AI as a productivity-enhancing tool. However, this convergence also exposes a limitation within the current literature. Much of the research explains why students use AI but provides comparatively less understanding of the conditions under which efficiency-oriented use remains educationally beneficial rather than becoming academically detrimental. Consequently, AI use should not be viewed simply as a technological behavior, but as a response to academic pressures, workload demands, and students' perceived need for learning support. Beyond academic efficiency, several studies identify psychological factors influencing AI adoption. Mental fatigue and academic stress encourage greater reliance on generative AI [9], while Johnston et al. [6] demonstrate that students with lower academic self-confidence are more likely to depend on AI. In contrast, students with greater confidence report using these technologies less and are more critical of others who use them. These findings collectively suggest that AI adoption cannot be explained solely by technological accessibility. Instead, individual learner characteristics, including confidence, stress, and cognitive demands, appear to shape patterns of AI dependence. Nevertheless, existing studies primarily establish associations between these variables, leaving unanswered questions regarding how these psychological factors interact with institutional environments and educational cultures.

The literature also presents a more nuanced understanding of students' attitudes toward generative AI than is often assumed. Polakova and Klimova [10] argue that Generation Z students' shorter attention spans and preference for rapid access to information make AI particularly attractive, while Johnston et al. [6] similarly suggest that impatience and outcome-oriented learning encourage technology use and dependence. However, these findings are partially contrasted by Chung [7], whose participants viewed generative AI primarily as a support tool rather than an unquestioned authority. This distinction indicates that students are not universally passive consumers of AI-generated information. Instead, many recognize its limitations and acknowledge the need for human judgment when evaluating AI outputs. Such differences across studies suggest that student use exists on a continuum ranging from supportive assistance to excessive dependence, rather than representing a single homogeneous pattern of behavior. The educational consequences of AI use likewise reveal a more complex picture. Abbas et al. [5] associate excessive use of ChatGPT with increased procrastination, memory loss, and poorer academic performance, while Alrasheed et al. [9] report declines in analytical reasoning, independent problem-solving, reflective thinking, and information evaluation among students who rely heavily on AI-generated responses. Taken together, these studies suggest that negative educational outcomes are linked less to AI use itself than to excessive or uncritical dependence on AI. This distinction represents an important conceptual point that is often overlooked within the literature. Rather than framing AI as inherently beneficial or harmful, existing evidence indicates that learning outcomes depend largely on how students engage with these technologies. The growing emphasis on AI literacy reinforces this interpretation. Alrasheed et al. [9] found that students possessing stronger AI literacy, including awareness of AI limitations, ethical considerations, fact-checking practices, and prompt evaluation, were significantly more capable of using AI critically rather than dependently. Sousa and Cardoso [8] similarly argue that AI literacy enables responsible and effective AI use. At the same time, the Digital Education Council [11] emphasizes that students must develop competencies in prompting and working effectively with Large Language Models. Collectively, these studies suggest that educational outcomes are influenced not only by the availability of AI but also by students' ability to critically evaluate AI-generated content. This shifts attention from restricting AI use towards developing the competencies necessary for responsible engagement. Institutional governance represents another recurring theme throughout the literature. Chung [7] reports that many students remain uncertain about institutional policies governing AI use despite acknowledging instances of academic misconduct, while Johnston et al. [6] found broad agreement among students that universities should establish clear guidance defining acceptable and unacceptable AI practices. These findings indicate that students' ethical concerns are not necessarily accompanied by institutional clarity. Instead, ambiguity surrounding acceptable AI use may contribute to inconsistent practices across higher education. The discussion of academic integrity reflects a similar pattern. Academic integrity is commonly defined as honesty, responsibility, and transparency in academic work [6], [12]. However, the widespread availability of generative AI has challenged traditional understandings of originality, as students increasingly rely on AI-generated material for assignments and express concerns about inaccurate information, fabricated references, and ethical misuse [7], [8]. Rather than suggesting that students intentionally engage in academic dishonesty, the literature indicates that many continue to use AI despite recognizing its limitations because they lack sufficient knowledge, confidence, or institutional guidance to use these tools appropriately [7]. Consequently, current evidence increasingly positions AI literacy and institutional policy as complementary rather than independent interventions. The Digital Education Council [11] extends this discussion by proposing practical assessment strategies, including the design of authentic assessments, clear AI guidance, staged assignments, oral verification, careful evaluation of referencing, and student accountability measures. Importantly, these recommendations move beyond detecting AI misuse towards redesigning learning environments that encourage critical engagement while reducing opportunities for inappropriate dependence. Collectively, the literature demonstrates broad agreement that higher education should shift from attempting to eliminate AI use towards managing it through assessment redesign, AI literacy development, and institutional policy. Despite the growing international literature, important gaps remain. Existing studies have primarily been conducted across diverse educational contexts, with limited evidence on

students' perceptions and practices in Kuwait. Educational policies, assessment cultures, institutional expectations, and students' experiences of AI adoption differ across national contexts, limiting the direct transferability of findings. Consequently, there remains insufficient understanding of how Kuwaiti university students perceive the educational benefits, risks, ethical challenges, and institutional expectations surrounding generative AI. Examining these issues within Kuwait, therefore, contributes context-specific evidence to an international literature that has largely overlooked the Gulf region, while also informing locally relevant institutional policies and educational practices.

III. STUDY IMPLEMENTATION

To address this study's four key questions, an online survey was administered to students enrolled in a private English-medium instruction university in Kuwait during the 2025-2026 academic year. A pilot survey was first administered to three classrooms, in which students provided feedback on the questions before the final form was submitted to the Institutional Review Board (IRB) for approval. The final version of the survey was then sent by email to all students enrolled in the university. These students ranged from those in the Foundations program (a pre-degree math and English literacy program) to those in the post-graduate level. Of those students, 418 responded. They answered an online form with 21 questions. The first part of the survey collected background information from the students, including their major, age, and self-rated level of experience with AI. The remainder of the survey included 16 self-evaluation questions that ranged from their use of AI to their understanding of ethics and academic integrity regarding its use. The full survey is outlined in Appendix A. Before analysis, it is important to note the limitations associated with this study. First, the findings are based on students from a single higher education institution, which may limit their generalizability to other universities in Kuwait or beyond. Second, the use of self-reported survey data introduces the potential for response and social desirability bias. Third, unequal subgroup sizes may have influenced comparisons across demographic groups. Finally, the cross-sectional design captures perceptions at a single point in time and therefore cannot determine changes in AI use or attitudes over time.

IV. RESULT ANALYSIS

The data collected from the student survey at the university (N=418), covering all 21 questions, were analyzed to understand the four key exploratory questions: normalization, use of the tool, ethical concerns, and expectations of guidance. As no factor is to be ruled out, each section of the survey is assessed to identify patterns in student stakeholders' perspectives on AI use throughout their educational journey.

1. Background

Of the students who responded (N=418), 55.7% were female, 42.1% were male, and the rest were undeclared. More than half of the participants (62.5%) were between the ages of 18 and 20, with the next largest group being 21-23. A small percentage (7.2%) was under 18, and an even smaller portion (2.6%) was over 26.

Table 1. Participants' Gender Distribution

Category	Numbers	Percentages
Female	233	55.7%
Male	176	42.1%
Undeclared	9	2.2%

Table 2. Participants' Age Distribution

Category	Numbers	Percentages
Under 18	30	7.2%
18-20	261	62.5%
21-23	100	23.9%
24-26	16	3.8%
AbLove 26	11	2.6%

Regarding their major, almost half of the participants were enrolled in computer science and engineering (22% each). Accounting was a close second with (12%). The rest spread relatively evenly across other majors offered at the university, including but not limited to business administration, English, finance, marketing, MIS, and mass communication (see Figure 1).

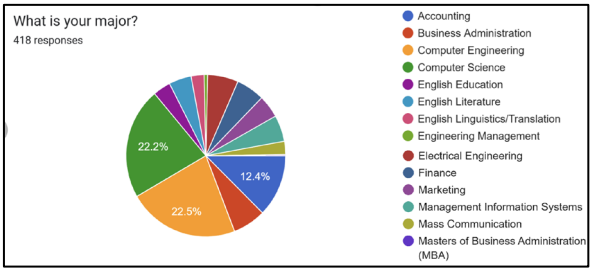


Figure 1. Majors

Regarding year of study, 20.1% were in Foundations, 30.6% were first-year, 19.1% were second-year, 14.6% were third-year, and 12% were fourth-year. A minority of participants were graduate students (3.6%).

Table 3. Participants' Year of Study Distribution

Category	Numbers	Percentages
Foundation	84	20.1%
Year 1	128	30.6%
Year 2	80	19.1%
Year 3	61	14.6%
Year 4	50	12%
Graduate Studies	15	3.6%

Students were asked to describe their proficiency with AI. While this question is difficult to pin down accurately, it offers a very poignant view of students' comfort and frequency of use [10]. The majority described themselves as beginners (31.8%) and intermediate (40%). 18.4% self-identified as advanced, which should be

considered a significant number for a relatively small pool of student participants addressing fairly new technology. A small percentage (9%) claimed they have never used AI.

2. Use of Generative AI in Academia

While completing assignments, 66.7% of students admitted to using generative AI. Among those who reported using AI, more than half (52%) reported regular use, with the rest reporting occasional or rare use. When asked for the specific purpose for their use of Generative AI in their academic work, Students were given the option to select multiple answers. Participants listed assistance with research as the major reason, at 75%. This was followed by help with assignments at 72.5%, although the extent of help or its definition was left vague. Many participants reported using AI for summarizing and editing (44% and 59%, respectively) and for solving math problems (40%). 25% reported using AI for coding assistance. A small percentage (12%) admitted to knowingly using AI to complete their assignments, thereby plagiarizing their work.

To further assess this, a Pearson Chi-Square test was conducted to determine whether there is an association between gender and AI use. Although more female participants responded affirmatively to using AI than male participants (see Figure 2), the difference was not statistically significant (p = .145).

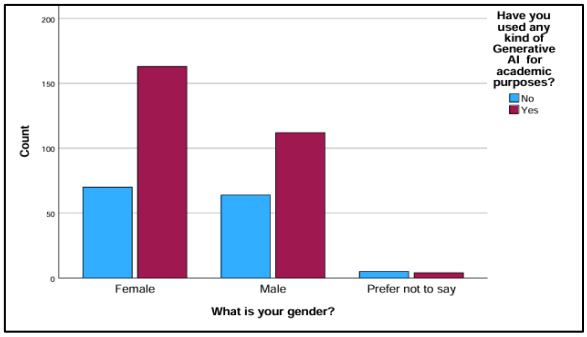


Figure 2. Gender and AI Use

Next, the test was conducted to determine any association between the major and the use of AI. Again, no statistical significance was noted (p=.794). However, an interesting result was observed in the percentage of use by major (see Figure 3). Of the 14 majors, 8 recorded around 65% affirmative use of AI for academic purposes: Accounting, Business Administration, Computer Engineering, Computer Science, Electrical Engineering, English Linguistics/Translation, Finance, and Management Information Systems. Another 4 majors recorded a 80% rate: English Education, English Literature, Marketing, and Mass Communication. Two majors recorded 100%; however, those were very small samples and can thus be neglected.

Table 4. Pearson Chi-Square tests involving the use of AI

Test Parameters	**Value**	**df**	**p-value**
Gender and Use of AI	3.866	2	.145
Major and Use of AI	8.711	13	.794
Year of Study and Use of AI	27.218	5	<.001*

*Statistically significant with p =0.0167 after Bonferroni correction.

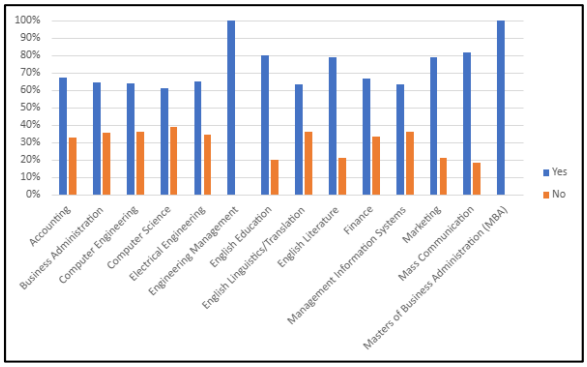


Figure 3. Major and AI Use Crosstabulation

This study also explored the relationship between year of study and AI use. Interestingly, a strong association was observed ($p < .001$; see Table 4). The Bonferroni correction was applied, yielding a new p-value of 0.0167. The results show that as students progress through college, the use of AI for academic purposes increases (see Figure 4). A Cramer's V test was implemented to determine the strength of this association. Cramer's V was 0.225, which indicates a moderately strong relationship. Further study is needed to determine the reasons for this increase. We cannot assess whether this change occurs as students progress, given the nature of our study. This level of analysis would require studying a cohort over the full course of their studies, not at a single point in time.

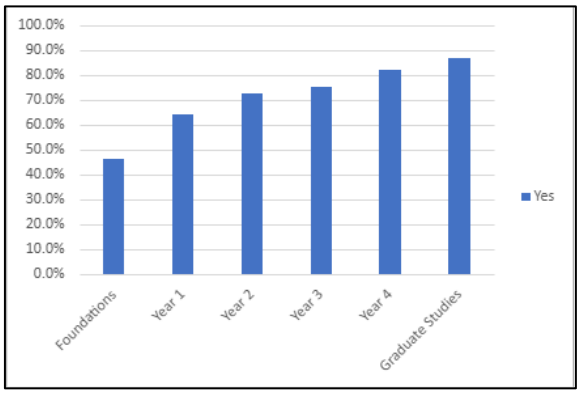


Figure 4. Year of Study and AI Use Crosstabulation

However, this sparked curiosity about how AI differs in the upper years of study. One question that needs exploration is how often they use AI. The Chi-Square test, with a Bonferroni correction (p-value = 0.0125), showed a significant association between year of study and the frequency with which participants use AI (see Table 5 and Figure 5). Interestingly, then, graduate students had the highest use of AI, with 40% reportedly using it several times per week (always). In contrast, foundation students had the lowest, with 5% using it several times per week (always).

Table 5. Pearson Chi-Square tests involving Year of Study

Test Parameters	**Value**	**df**	**p-value**
How often AI is used and Year of Study	50.828	25	.002*

Understanding output before use and Year of Study	25.185	20	.194
Need for guidelines and Year of Study	12.548	10	.250
Using AI when not allowed and Year of Study	12.182	5	.032

*Statistically significant with p =0.0125 after Bonferroni correction.

The study also investigated an association between year of study and understanding of AI output before use, but no statistically significant association was found (see Table 5). There was also no statistically significant association between year of study and their need for guidelines for using AI (see Table 5).

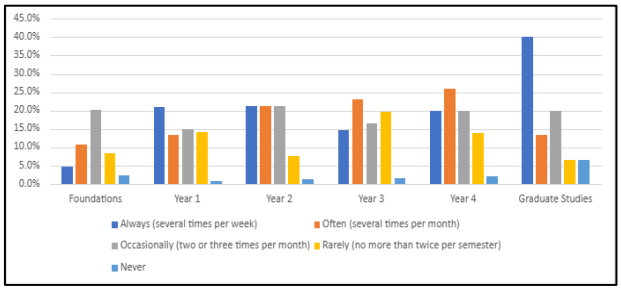


Figure 5. Year and Frequency of AI Use Crosstabulation

Other responses included using AI as a tutor, translator, and for daily life tasks. Active users were then prompted to rate how helpful they find AI to be in completing their assignments. More than half (64%) found it helpful, while only 3% found it unhelpful. To further understand and analyze their use of AI, students were asked whether they copied and pasted the answers generated by AI or put in some effort to understand the generated answers and then form their own response. The purpose here was to assess whether AI was being used as a tool for help or as a crutch. Fortunately, a small percentage (4%) reported copying and pasting, and more than half (60%) put in some effort to understand. The level of effort ranged from little to full understanding, and this gap is a key area of interest for further research aimed at improving the use of AI as an academic tool.

3. Ethics and Academic Integrity

The ethical implications of AI tools in academia were also explored in this study. Participants were asked several questions to gauge their moral stance and understanding of AI use in their work. They were asked whether completing assignments with AI support was ethical. While almost half said 'Yes' (40.7%) and only 21.5% said 'No', a substantial number were unsure (37.8%). This indicates a lack of clarity and understanding regarding the ethical use of AI in classwork. This further indicates an urgent need for guidance and regulations as students continue to use AI.

This need for guidance became more apparent when participants were asked whether they had conversations with their professors about the ethical use of AI. For example, have their professors warned them about using AI to complete their assignments? The responses were split with 51.4% responding 'Yes' and 48.6% responding 'No'. To further delve into this ethical issue, participants were asked whether, if the rules did not allow the use of AI, they would use it anyway. More than half (65%) responded that they would not. Furthermore, they were asked if they had ever faced consequences for using AI in their assignments. The majority answered negatively (80%). While some hypotheticals were used to gauge current understanding, this

does not negate the deduction that there exists a need for faculty guidance and a faculty role in adopting or guiding AI use in the classroom. Fortunately for academia, the question of whether they would prefer a clear policy on the use of AI in academic work saw more than half (56.9%) respond affirmatively. Preferring a clear policy from their professors shows that students are not lazy or unaware; they are simply growing up in a digital age and need guidance on what that means for their education. A Pearson Chi-square test was conducted to examine participants' willingness to use AI despite a clear policy against it. While hypothetical, this frames students' existing stance on plagiarism. The Chi-square test was conducted to examine the association between the year of study and the use of AI, despite it being prohibited. No statistically significant association was found (see Table 5). With that said, a high percentage of participants in the later stages of the study reported being willing to ignore policies prohibiting the use of AI (see Figure 6). Further study is required to understand the reasons for this perspective.

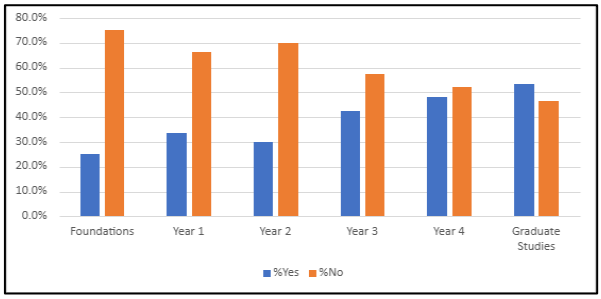


Figure 6. Year and Understanding of Ethicality Crosstabulation

The study did, however, show that students received some guidelines about AI use from their professors, but they were specific about the tool being ChatGPT. It also showed that ChatGPT is the most popular tool, with 69% listing it as their go-to AI tool. This led to the conclusion that policies set by professors should not be exclusive and need to include most, if not all, generative AI tools. The next question indicated that participants understood the need for guidelines regarding the use of AI tools, such as ChatGPT. When asked whether they believe ChatGPT should be integrated into the education system as a learning tool, 64% responded 'yes, with clear guidelines.' This overwhelming affirmative response emphasizes that students want AI as a tool and indicates that AI is a fixture in academic settings. Their awareness of the addition of guidelines reflects a clear understanding of the need on both sides – students and professors – to adopt AI ethically.

V. CONCLUSION

The close ratio of male and female participants allows for a relatively equal interpretation of the data across genders. This study confirms what is commonly known among faculty: students are using Generative AI to complete their academic work. It is no surprise that students find AI helpful, as that is a core aspect of AI's functionality. However, the study highlighted that their use of AI has not extinguished their desire to learn. A surprising result, though, is the increase in AI use among foundation-year students to graduate students. One would assume that students become more independent learners as they progress through their studies. Of course, this study cannot be taken as a progression and is addressed as a snapshot. This study shows higher reported AI use among students in later years of study, and that remains a point for further exploration. Given the cross-sectional design, these findings should not be interpreted as evidence of increasing dependence on AI over time. Developing students' understanding of the capabilities and limitations of generative AI may encourage more informed and responsible use of these tools. This was echoed in Tian et al. [9], where a strong understanding

of the limitations led to less dependability. Further research is warranted to explore the factors associated with the higher levels of reported AI use among students at different stages of their academic studies.

The ethicality of AI use is always in debate [13]. The results of this research do indicate a need and desire among university students for clarity in this matter. This is in line with the results of Johnston et al. [6], Chung [7], and Sousa & Cardoso [8]. However, the findings indicate that students continue to use AI for academic purposes while also expressing concerns regarding its ethical use. Although AI use remains common, the present study did not directly examine students' motivations for these behaviors or the extent to which institutional policies influenced their decisions. This is clear from the data here, where students have ignored AI policies and continued to use it in their academic work. The findings suggest that reported AI use differs across years of study, highlighting the need for further investigation into how students interpret and respond to institutional AI policies. Further study is also required to understand such behavior and the factors that drive older and more mature students to do so. On the other hand, the study also showed that it is unclear to students what constitutes ethical use and what constitutes plagiarism when it comes to AI. This ambiguity highlights the need for clearer institutional guidance and policies regarding acceptable AI use in academic work. Furthermore, the need for the policy to be inclusive of most AI tools, rather than focusing on one popular tool at the moment, is important. Students need an understanding of AI's capacity for academic work.

A noteworthy aspect of this study was students' awareness that AI can be a helpful tool when used correctly and ethically. Some have stated that they use AI tools as tutors to help with understanding the material. Others have used AI tools as a support for completing academic work, not as a source to copy and paste. Rather than prohibiting generative AI outright, educators may benefit from adopting clear expectations and guidance that promote responsible and ethical AI use:

- Including a clear policy on the use of AI in the syllabus.
- Ensuring students understand what is considered the use of AI
- Guiding students on how to source their use of AI in their academic work clearly

VI. FUTURE WORK

We hope to understand AI use in academia from all perspectives, and to eventually provide holistic, ethical suggestions that consider all those involved in the process. Further study is required to investigate the factors driving the increase in AI use among students, from those starting college to those nearing graduation. A longitudinal study might help investigate whether there is an increase in dependence on AI as students progress through their study years. The need to investigate current academic policies on the use of AI and the reasons students disregard them is also on our radar. This paper is the beginning of a larger project exploring AI in academia. Thus, our next step is to assess faculty uses of AI, including their approach to student use of AI in their classrooms.

REFERENCES

[1]. Matthews, K. E., & Dollinger, M. (2023). Student voice in higher education: The importance of distinguishing student representation and student partnership. Higher Education, 85(3), 555–570. https://doi.org/10.1007/s10734-022-00851-7

[2]. Roberts, D. M. (1980). The impact of electronic calculators on educational performance. Review of Educational Research, 50(1), 71–98. https://doi.org/10.3102/00346543050001071

[3]. Mogari, D., & Faleye, S. (2012). Introducing calculators to learners early in their schooling: The effect on long-term arithmetic proficiency. African Journal of Research in Mathematics, Science and Technology Education, 16(3), 363–375. https://doi.org/10.1080/10288457.2012.10740750
[4]. Granić, A. (2022). Educational technology adoption: A systematic review. Education and Information Technologies, 27(7), 9725–9744. https://doi.org/10.1007/s10639-022-10951-7
[5]. Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. International Journal of Educational Technology in Higher Education, 21(1), 1–22. https://doi.org/10.1186/s41239-024-00444-7
[6]. Johnston, H., Anderson, M., & Lee, S. (2024). Student perspectives on the use of generative artificial intelligence in higher education. International Journal of Artificial Intelligence in Education, 34(2), 245–263. https://doi.org/10.1007/s40593-024-00341-8
[7]. Chung, J. (2026). The use and usefulness of generative AI in higher education. Smart Learning Environments, 13(1), 1–19. https://doi.org/10.1186/s40561-026-00315-9
[8]. Sousa, A. E., & Cardoso, P. (2025). Use of generative AI by higher education students. Electronics, 14(7), Article 1258. https://doi.org/10.3390/electronics14071258
[9]. Tian, J., Zhang, Y., Liu, F., Chen, X., & Wang, H. (2025). Learners' dependence on AI and critical thinking: The psychological mechanism of fatigue and the social buffering role of AI literacy. Acta Psychologica. https://doi.org/10.1016/j.actpsy.2025.105420
[10]. Polakova, B., & Klimova, B. (2019). Mobile technology and Generation Z in the English language classroom: A preliminary study. Education Sciences, 9(3), Article 203. https://doi.org/10.3390/educsci9030203
[11]. Digital Education Council. (2026). The impact of generative AI on student engagement and ethics in higher education. Journal of Educational Technology Systems, 55(1), 44–61. https://doi.org/10.1177/00472395261234567
[12]. East, J., & Donnelly, L. (2012). Taking responsibility for academic integrity: A collaborative teaching and learning design. Journal of University Teaching & Learning Practice, 9(3), Article 2. https://doi.org/10.53761/1.9.3.2
[13]. Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence, 5, Article 100152. https://doi.org/10.1016/j.caeai.2023.100152

APPENDIX A

Link to survey: https://forms.gle/CECHMMeouowH2sCY8

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Analysis Of Existing Approaches In The Field Of Cryptographic Protocols In Uavs

Ilaha Gahramanova
Azerbaijan Technical University
Department of Cybersecurity
Azerbaijan, Baku
ilaha.qahramanova@aztu.edu.az
ORCID:0009-0007-9761-1922

Abstract
Since uncrewed aerial vehicles are widely used in military, industrial, logistics, and civil sectors, providing them with secure communication channels is of strategic importance. There is a serious need to develop energy-efficient, computationally lightweight authentication algorithms and reliable cryptographic key exchange protocols, which are the basis of secure communication. The scientific study of functional architecture and data exchange models plays a key role in ensuring the system's reliability, scalability, and cybersecurity. This review article analyzes the role of existing authentication protocols, lightweight cryptographic algorithms, PUF-based authentication, and blockchain technologies in UAV security. Future research directions - post-quantum cryptography and security mechanisms integrated with artificial intelligence - are also discussed.

Keywords: UAV, Elliptic Curve Cryptography, lightweight cryptographic protocols, post-quantum cryptography, blockchain technologies.

I. INTRODUCTION

In recent years, various lightweight cryptographic protocols have been proposed for UAV environments. These protocols are based on technologies such as symmetric and asymmetric cryptography, elliptic curve cryptography (ECC), hash-based mechanisms, public key infrastructure (PKI), lightweight authentication approaches, and blockchain and physical-layer-based methods. However, there are significant differences among existing approaches in security level, computational complexity, energy consumption, latency, and scalability.

Research Objective

The primary objective of this systematic literature review is to identify, analyze, and compare recent lightweight cryptographic authentication protocols developed for Unmanned Aerial Vehicle (UAV) networks and the Internet of Drones (IoD). The review also aims to identify existing security challenges and emerging research directions, including a comprehensive review and classification of lightweight cryptographic authentication protocols for UAVs, analysis of existing approaches, highlighting their strengths, limitations, and research gaps, and an overview of key challenges and future research directions for secure and efficient UAV authentication.

To achieve these objectives, the following research questions (RQs) were formulated:

- **RQ1:** What lightweight authentication protocols have been proposed for UAV and Internet of Drones environments?
- **RQ2:** Which cryptographic techniques are most frequently employed in these protocols?
- **RQ3:** What security properties and attack resistance are provided by existing authentication schemes?
- **RQ4:** What are the current limitations and future research directions in UAV authentication?

II. RESEARCH METHODOLOGY

The literature search was conducted using the following electronic databases: IEEE Xplore Digital Library, Scopus, Web of Science, ScienceDirect (Elsevier), SpringerLink, MDPI, and ACM Digital Library. Publications from January 2018 to May 2025 were considered. The search was performed using combinations of the following keywords: "UAV Authentication", "Drone Authentication", "Internet of Drones", "IoD Security", "Lightweight Authentication", "Mutual Authentication", "Authentication Protocol", "Lightweight Cryptography", "Elliptic Curve Cryptography (ECC)", "PUF Authentication", "Blockchain Security", "Post-Quantum Cryptography", "Firmware Security", and "Drone Communication Security". The selected studies were screened by title, abstract, and full-text analysis. Relevant information, including the authentication method, cryptographic techniques, security features, and performance characteristics, was extracted and comparatively analyzed to identify current research trends, limitations, and future directions.

III. ATTACKS ON UAV SYSTEMS

Uncrewed aerial vehicles (UAVs), or drones, provide numerous real-time services, from intelligent transportation systems and military applications to emergency rescue and environmental monitoring [1]. The low cost and flexible operational capabilities of these devices are the main reasons for their popularity [2]. However, the exchange of data over open communication channels [3] and the frequent deployment of UAVs in dangerous or uncontrolled areas make them vulnerable to several security attacks (Table 1).

Security issues are mainly divided into the following categories:

1) Network-based attacks: In the event of a physical capture of a UAV, confidential security parameters stored in its memory can be stolen, leading to privacy and security breaches for the entire system [3].

2) Physical capture attacks: Messages transmitted between UAVs or between UAVs and a ground station are subject to impersonation, session key disclosure, message replay, man-in-the-middle (MitM), monitoring, and eavesdropping attacks [4].

IV. CRYPTOGRAPHIC PROTECTION MECHANISMS IN UAV SYSTEMS

The UAV system is built on a multilayer architecture that requires real-time coordination. The main security requirements in these systems are as follows:

- Authentication and identification – mutual recognition between the UAV and the ground station.
- Confidentiality and information secrecy – encryption of flight coordinates, sensor data, and command packets.
- Integrity – prevention of data modification.
- Lightweight computing – cryptography suitable for limited energy and processor power.
- Physical security – resistance to device capture and reverse engineering attacks.

These requirements form the main design motives of various protocols.

The difference in the threat model for UAVs is that:

- Compromise can lead to real physical incidents.
- Resource limitations (CPU, energy) make it difficult to implement classical cryptographic protocols.
- Mobile nodes are sensitive to round-trip delay.

As UAVs are distributed and resource-constrained systems, they are exposed to various levels of security threats. These threats primarily affect communication, navigation, cryptographic mechanisms, software, the physical environment, and resource utilization (Table 1). As shown in Fig.1, research on UAV cybersecurity has grown steadily between 2019 and 2026, reflecting increasing scientific interest in this field. The number of publications rose gradually from 2019 to 2022, followed by a more pronounced increase during 2023–2025, indicating the rapid development of UAV cybersecurity research. Although the number of publications indexed for 2026 is based solely on records available at the time of data collection and does not reflect the full annual output, the results confirm that this topic continues to attract significant research attention. Overall, the publication trend highlights the growing importance of developing effective cybersecurity solutions for UAV systems.

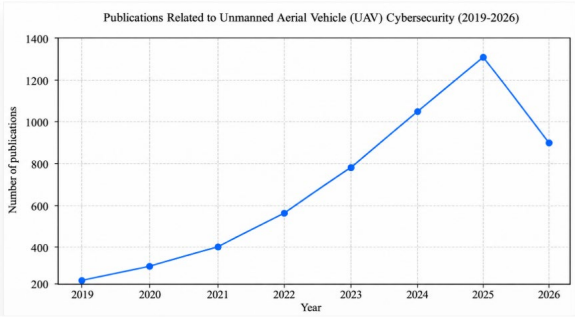


Figure 1. Publications related to Unmanned Aerial Vehicle cybersecurity in 2019-2026

Replay and Man-in-the-Middle (MITM) attacks can lead to the falsification of control commands and the alteration of telemetry data. MAC/HMAC mechanisms ensure message integrity. Nonce and timestamp-based authentication, session identifiers, mutual authentication, and ECDH-based key agreement mechanisms are used to prevent such attacks. GPS spoofing and GPS jamming attacks can cause the UAV to be redirected to incorrect coordinates or to lose navigation capabilities. To mitigate these risks, multi-sensor fusion (GPS and IMU integration), signal anomaly analysis, anti-jamming antennas, and inertial navigation mode-switching mechanisms are employed. In the field of cryptographic security, key extraction attacks can lead to the compromise of long-term secret keys and complete loss of control over the system. Approaches such as Secure Element (a specialized hardware component in cryptographic security, where secret keys and critical cryptographic operations are physically and logically protected), key generation based on physically unclonable functions (PUFs), and regular key renewal are used to mitigate these threats. Firmware injection attacks allow the execution of malicious code. The Secure Boot mechanism and the verification of software digital signatures minimize this risk. In terms of physical security, reverse engineering and the extraction of cryptographic keys are possible following device capture. Tamper-resistant hardware solutions and automatic memory erasure mechanisms are used against this. In a device takeover attack, it is possible to extract cryptographic keys through reverse engineering, modify sensor and control data, miscalculate the flight trajectory, and execute control decisions incorrectly, thereby completely disrupting the mission, causing the UAV to go out of control, or causing it to crash. Resource exploitation and DoS attacks cause the UAV to rapidly exhaust its power resources and disrupt real-time operations. To mitigate the impact of these attacks, lightweight cryptographic algorithms, traffic filtering, and rate-limiting techniques are used.

In identity-spoofing attacks, a rogue ground station can control the UAV. This threat is eliminated by certificate-based authentication and mutual authentication protocols. At the same time, hash functions, MAC mechanisms, and sequence numbers are used to prevent data integrity attacks during packet modification (Table 1). Consequently, the security of UAV systems requires a multi-level and integrated defense approach. This approach, based on the Defense-in-Depth principle, significantly increases the system's resilience and reliability against various threat vectors. For the cybersecurity of the UAV system, lightweight cryptographic protocols, certificate-based authentication, mutual identification, key exchange protocols, hash functions, and MAC/HMAC are the most powerful defense mechanisms against many types of attacks (Table 1).

Table 1. Attack–Defense mechanism analysis for Unmanned Aerial Vehicle systems

Type of attack	Defense mechanism
Replay attack	Nonce and timestamp-based authentication, session IDs, challenge–response mechanism
MITM (Man-in-the-Middle)	Mutual authentication, key agreement (ECDH), MAC / HMAC usage
GPS spoofing	Multi-sensor fusion (GPS + IMU), signal strength, and anomaly analysis
GPS jamming	Anti-jamming antennas, switching to inertial navigation mode
Key extraction attack	Secure element, PUF-based key generation, frequent key updates
Firmware injection	Secure boot, verification of firmware digital signature
Device hijacking	Tamper-resistant hardware, automatic memory wipe, and ampere-resistant hardware, automatic memory deletion
DoS/power depletion attack	Lightweight cryptography, traffic filtering, and rate-limiting
Impersonation attack	Certificate-based authentication, mutual identification protocols

V. CURRENT APPROACHES IN THE FIELD OF LIGHTWEIGHT CRYPTOGRAPHIC PROTECTION MECHANISMS IN UAVS

Researchers have developed various lightweight cryptographic protocols to address the aforementioned security problems and provide suitable solutions for resource-constrained UAVs. These protocols are based on the following technologies:

1. Lightweight authentication and key exchange protocols

One of the most widely used methods in UAV communication is the combination of symmetric key and ECC. By ensuring the secure exchange of symmetric keys through ECC, the computational load is reduced, and high-speed data transmission is protected [1]. This approach is also suitable for real-time encrypted data transmission during flight [1].
ECC is the most widely used asymmetric cryptography mechanism in UAV networks because it provides high security with short keys. An ECC- and symmetric-key-based protocol is presented [1-3].
A new authentication and key agreement scheme between the user and the accessed PUA, with the help of the server, is presented [4]. After successful mutual authentication between the user and the PUA, the established session key enables them to communicate securely, preventing various known attacks by the adversary [4].
A lightweight scheme for remote user authentication and key agreement in the IoT environment has been proposed [4]. This protocol minimizes computational costs and provides mutual authentication. In the same

direction, they have succeeded in protecting identity confidentiality using the temporary credential-based anonymous authentication (TCALAS) system [5].
The choice of ECC is determined by its relatively shorter key size compared to other asymmetric encryption algorithms, such as RSA (Rivest-Shamir-Adleman). After successful authentication, both parties generate a secret session key for secure communication [5]. To address this important problem in the UAV system, a new time- and credential-based anonymous lightweight user authentication mechanism, called TCALAS, has been developed [5]. *A3*, *A5*, and *A6* are not fully considered in the article.
TCALAS has been improved because it is weak against tracking and stolen-authenticator attacks, and because it cannot be scaled. To overcome security problems with lower time and resource consumption, a temporary credential-based anonymous lightweight authentication scheme (iTCALAS) has been proposed for UAV networks [6]. The improved scheme, which uses lightweight symmetric-key primitives and temporary credentials, provides security against many known attacks, including tracking and stolen authenticators [6]. Security measures for attack types such as *A3*, *A4*, *A5*, and *A6* are not considered.
The PASKE-IoD protocol combines both confidentiality and authentication, making the data exchange more secure [7]. Vulnerabilities are observed against attack types such as *A3*, *A5*, *A6* [4-7].

2. PUF-based security

PUF technology has emerged as an effective approach to UAV security in recent years. A PUF uses the physical properties of semiconductor devices to create a unique, difficult-to-clone "digital fingerprint" for each device. The CoMSeC++ protocol provides unique device identification using PUFs and provides an attack-resistant authentication system [8]. The same principles have been applied in the PLAKE protocol, preventing the forgery of key exchanges [9]. The proposed protocols use PUFs to prevent the leakage of private keys and provide robust protection even in the event of physical capture [8], [9]. Vulnerabilities to attacks such as A3 and *A5* [8], [9] are observed. An S-IoD framework is presented for the UAV environment that independently collects sensitive information [10]. A lightweight privacy-preserving scheme (L-PPS) is introduced to reduce the computational cost of the authentication protocol. L-PPS is designed to ensure reliability between IoT devices during authentication [10]. Although many threat scenarios are considered during protocol development, it remains vulnerable to attacks such as *A3*, *A5*, and *A6*. A lightweight, privacy-preserving mutual authentication and key agreement protocol, PMAP, is proposed [11]. It uses PUF and a chaotic system to support mutual authentication and generate secure session keys between communicating entities in an IoD system. PMAP consists of two schemes: PMAPD2Z (which mutually authenticates the PUA and the zone service provider and generates secure session keys) and PMAPD2D (which mutually authenticates the PUAs and generates secure session keys) [11]. The proposed protocol achieves confidentiality, integrity, mutual authentication, non-repudiation, perfect forward secrecy (PFS), perfect backward secrecy (PBS), defense against Denial-of-Service (DoS) attacks, man-in-the-middle (MITM) protection, and Drone-to-Drone (D2D) security [12]. The paper proposes a secure and lightweight biometric-based authenticated key agreement (AKA) protocol for the Internet of Drones (IoD) environment, utilizing lightweight cryptographic operations such as secure bitwise XOR operations, one-way hash functions, and the user's biometric verification for mutual authentication. A secure and lightweight biometric-based authenticated key agreement (AKA) protocol for the Internet of Drones (IoD) environment, utilizing lightweight cryptographic operations such as secure bitwise XOR operations, one-way hash functions, and users' biometric verification for mutual authentication [13]. Vulnerabilities to attacks such as A3 and *A5* are observed.

3. Blockchain and Distributed Trust Mechanisms

This approach allows for the implementation of decentralized trust systems in drone control. Blockchain is used to protect data integrity, facilitate secure key exchange, and track decisions in drone networks [14], [15]. Vulnerabilities are observed against attack types such as *A3*, *A5*, and *A6*.
A possible solution is proposed for each logical layer of the IoT architecture [15]. In addition, encryption mechanisms, privacy decentralization, secure communication, lightweight cryptographic protocols, and solutions against many security threats have been developed [15].

4. Artificial Intelligence and Deep Learning-Based Security

In recent studies, the combination of deep learning models with post-quantum cryptography is considered a promising direction for future drone defense systems [16]. This approach allows for the development of real-time attack detection and adaptive response mechanisms.
A review of approaches in the field of cryptographic protocols for UAVs shows that many research gaps remain and that the scope of development directions needs to be expanded.
In terms of energy optimization, most UAVs have limited battery life. Therefore, the energy optimization of cryptographic operations remains an open problem—particularly for integration with 6G and massive IoT. Future UAVs will operate in conjunction with 6G networks and require multiple low-latency connections, as well as protection against artificial intelligence-based attacks. The cyber-physical nature of UAVs necessitates protection against machine-learning-based attacks [16].

Table 2. Comparison of Unmanned Aerial Vehicles' security protocols

№	Protocol / Research	Cryptographic mechanism	Advantages (brief)	Disadvantages (brief)	Verification / Analysis Tool
1	Nyangaresi et al., 2023	Symmetric AES + ECC	Lightweight, high-speed	ECC operations are still relatively heavy for UAVs A3, A5, and *A6*.	AVISPA, BAN-logic
2	Yahuza et al., 2021	Hash-based + ECC	Edge-assisted, low-latency	Dependence on boundary structure. *A3*, *A5, A6*.	Scyther, ProVerif
3	Tian et al., 2024	ECC + Physical security	Resistance to physical capture	Higher computational load. Weak DoS optimization.	AVISPA, RoR model
4	Wazid et al., 2018	Hash chain + ECC operations against	DoS and MITM.	ECC operations. *A3*, *A5, A6*.	AVISPA, BAN-logic
5	Srinivas et al., 2019 (TCALAS)	Temporal Credentials + Hash	Anonymity, very light	Credential management is difficult. *A3*, *A5, A6*.	AVISPA
6	Ali et al., 2020	Lightweight hash + XOR	Minimal overhead for Smart-city	Formal proof is weak. *A3*, *A4 A5, A6*.	AVISPA, BAN-logic
7	Tanveer et al., 2021 (PASKE-IoD)	ECC + Pseudonym	Privacy strong	*A3*, *A5, A6*.	ProVerif, AVISPA

8	Mall et al., 2021 (CoMSeC++)	PUF + Hash	High physical security with PUF	*A3, A5.*	AVISPA, PUF reliability tests
9	Roy et al., 2022 (PLAKE)	PUF + Hash	PUF is resistant to attacks	*A3, A5.*	AVISPA, Formal PUF modeling
10	Deebak & Al-Turjman, 2020	Lightweight crypto + Privacy masking	Privacy is preserved	Computational burden can be high in some PUAs. *A3*, *A5, A6.*	AVISPA
11	Pu et al., 2022	ECC + Mutual Authentication	Drone-Drone and Drone-Server exchange strong	ECC workload. *A3*, *A5, A6.*	Scyther, AVISPA
12	Ko et al., 2021	Symmetric crypto + OTP	High security for the military zone	OTP management	AVISPA, Simulation-based proof
13	Akram et al., 2021	Hash-based + ECC Lightweight	Smart-city video transmission is insecure	*A3, A5.*	ProVerif, BAN-logic
14	Naveen et al., 2025	Blockchain + Lightweight cryptography	Full immutability, suitable for IoT	Blockchain latency. *A3, A6.*	Hyperledger testbed, Security analysis
15	Chowdhury et al., 2022	Blockchain + Hash	Decentralized authentication	High computation and latency. High time consumption. *A3, A6.*	AVISPA, ProVerif
16	Sharma & Rani, 2025	Post-Quantum Cryptography + ML	Future defense with PQC	PQC is too heavy and problematic for UAVs. *A3, A6.*	Security proofs, PQ-hardness assumptions

VI. CRITICAL ANALYSIS OF RESEARCH GAPS AND UPDATES

To ensure secure communication through UAV authentication models and enable successful, reliable flight missions in UAV systems, numerous authentication mechanisms have been proposed in the literature [1-16]. Using sixteen papers, the security of UAV authentication is analyzed. The analysis shows that most UAV authentication schemes are lightweight in terms of communication costs (Table 2). However, storage costs or energy consumption are not reported in many authentication studies. Formal models widely used in security analysis are identified, while most studies do not cover many attacks targeting UAV systems (Table 3).

Table 3. Comparison of security features

S/N	**Reference**	**A 1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**
1	Nyangaresi et al. (2023)										

2 Yahuza et al. (2021)

RQ3: Security Properties and Attack Resistance

Most protocols provide mutual authentication, confidentiality, integrity, forward secrecy, and resistance against replay, impersonation, and man-in-the-middle attacks. Their security is commonly validated using ROM, RoR, BAN logic, and the Dolev–Yao model. However, UAV-specific threats such as GNSS spoofing, firmware attacks, physical capture, and side-channel attacks are rarely considered.

Key Finding: Existing schemes effectively resist traditional network attacks but provide limited protection against UAV-specific threats.

RQ4: Limitations and Future Directions

Current protocols are largely validated through theoretical analysis, with limited real-world UAV implementation and testing. Future research should focus on scalable authentication for drone swarms, dynamic UAV networks, secure firmware verification, and hybrid, lightweight cryptographic frameworks that adapt to emerging cyber threats.

Key Finding: Future authentication protocols should improve scalability, adaptability, and resilience against both conventional and UAV-specific attacks.

VIII. CONCLUSION

Key agreement and lightweight cryptographic security protocols for UAVs are a rapidly developing research area. Although current approaches provide a certain level of security, formal verification, energy optimization, and the application of post-quantum cryptography still require further research. This review systematically compares state-of-the-art lightweight cryptographic protocols for UAVs and identifies future research directions. Researchers have found partial solutions to these problems with lightweight cryptography and authentication protocols such as ECC, PUF, and blockchain. However, the rise of quantum computing poses new threats to existing cryptographic systems. Therefore, future research should focus on implementing PQC and developing hybrid authentication models to ensure secure communication between classical and quantum UAV networks. This approach will ensure the long-term cybersecurity and resilience of UAV systems.

REFERENCES

[1]. V. O. Nyangaresi, H. M. Jasim, K. A. A. Mutlaq, Z. A. Abduljabbar, J. Ma, I. Q. Abduljaleel, and D. G. Honi, "A symmetric key and elliptic curve cryptography-based protocol for message encryption in unmanned aerial vehicles," Electronics, vol. 12, no. 17, Art. no. 3688, 2023, doi: 10.3390/electronics12173688.

[2]. Yahuza, M., Idris, M. Y. I., Wahab, A. W. A., Nandy, T., Ahmedy, I. B., & Ramli, R. (2021). An edge-assisted secure lightweight authentication technique for safe communication on the Internet of Drones Network. *IEEE Access*, *9*, 31420-31440.

[3]. Tian, C., Ma, J., Li, T., Zhang, J., Ma, C., & Xi, N. (2024). A provably and physically secure UAV-assisted authentication protocol for IoT devices in unattended settings. *IEEE Transactions on Information Forensics and Security*, *19*, 4448-4463.

[4]. Wazid, M., Das, A. K., Kumar, N., Vasilakos, A. V., & Rodrigues, J. J. (2018). Design and analysis of a secure lightweight remote user authentication and key agreement scheme in the Internet of Drones deployment. *IEEE Internet of Things Journal*, *6*(2), 3572-3584.

[5]. Srinivas, J., Das, A. K., Kumar, N., & Rodrigues, J. J. (2019). TCALAS: Temporal credential-based anonymous lightweight authentication scheme for the Internet of Drones environment. *IEEE Transactions on Vehicular Technology*, *68*(7), 6903-6916.

[6]. Ali, Z., Chaudhry, S. A., Ramzan, M. S., & Al-Turjman, F. (2020). Securing smart city surveillance: A lightweight authentication mechanism for uncrewed vehicles. *IEEE Access*, *8*, 43711-43724.

[7]. Tanveer, M., Shah, H., Chaudhry, S. A., & Naushad, A. (2021). PASKE-IoD: Privacy-protecting authenticated key establishment for Internet of Drones. *IEEE Access*, *9*, 145683-145698.

[8]. Mall, P., Amin, R., Obaidat, M. S., & Hsiao, K. F. (2021). CoMSeC++: PUF-based secured light-weight mutual authentication protocol for Drone-enabled WSN. *Computer Networks*, *199*, 108476.

[9]. Roy, S., Das, D., Mondal, A., Mahalat, M. H., Sen, B., & Sikdar, B. (2022). PLAKE: PUF-based secure lightweight authentication and key exchange protocol for IoT. *IEEE Internet of Things Journal*, *10*(10), 8547-8559.

[10]. Deebak, B. D., & Al-Turjman, F. (2020). A smart, lightweight privacy-preservation scheme for IoT-based UAV communication systems. *Computer Communications*, *162*, 102-117.

[11]. Pu, C., Wall, A., Choo, K. K. R., Ahmed, I., & Lim, S. (2022). A lightweight, privacy-preserving mutual authentication and key agreement protocol for the Internet of Drones environment. *IEEE Internet of Things Journal*, *9*(12), 9918-9933.

[12]. Ko, Y., Kim, J., Duguma, D. G., Astillo, P. V., You, I., & Pau, G. (2021). Drone secure communication protocol for future sensitive applications in a military zone. *Sensors*, *21*(6), 2057.

[13]. Akram, M. W., Bashir, A. K., Shamshad, S., Saleem, M. A., AlZubi, A. A., Chaudhry, S. A., ... & Zikria, Y. B. (2021). A secure and lightweight drone-access protocol for smart city surveillance. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 19634-19643.

[14]. Naveen, V., Narang, S., Ramesh, P. S., Ponmalar, A., & Sudha, I. (2025, April). Secure Data Transmission in IoT Networks Using Blockchain and Lightweight Cryptography: Safeguarding the Connected Future. In *2025, the 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)* (Vol. 3, pp. 537-542). IEEE.

[15]. Chowdhury, N. J. K., Alam, K. M. R., & Islam, M. M. (2022, April). Security and Privacy in IoT using Blockchain and Lightweight Cryptographic Protocol. In *2022, IEEE 7th International Conference for Convergence in Technology (I2CT)* (pp. 1-5). IEEE.

[16]. Sharma, A., & Rani, S. (2025). Post-quantum cryptography (PQC) for IoT-consumer electronics devices integrated with deep learning. *IEEE Transactions on Consumer Electronics*, *71*(2), 4925-493

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

SHRUMAC: An IoT-Smart Hygiene, Refuse Monitoring and Alert Controller System for Real-Time Institutional Campus Sanitation Management

Shobowale, Rukayat Temitope
Department of Computer Science and Informatics,
Faculty of Computing
Federal University Otuoke
Bayelsa State, Nigeria.
shobowalett@fuotuoke.edu.ng
ORCID: 0009-0009-4629-831X

Maclean, Friday Maclean
Department of Computer Science and Informatics,
Faculty of Computing
Federal University Otuoke
Bayelsa State, Nigeria

Abstract
Managing waste disposal in congested areas such as higher institutions, markets, and worship centers is becoming a serious concern; untimely disposal of this refuse poses a great challenge to human health and the environment at large. This research harnesses the strengths of smart technology to develop SHRUMAC: A Smart Hygiene, Refuse Monitoring, and Alert Controller System for Real-Time Institutional Campus Sanitation Management. The SHRUMAC uses ultrasonic sensors to monitor waste bin levels and Bluetooth to send alerts to a mobile app called the Technology-Enabled Monitoring and Alert Controller (TEMAC App). When the SHRUMAC is full, it automatically notifies designated personnel (cleaners), ensuring timely waste disposal. It also features an automatic lid operation for hands-free use, improving hygiene. The SHRUMAC system helps manage waste more efficiently by reducing overflow and creating a cleaner campus environment in our institutions. The mobile app is built with MIT App Inventor and powered by a rechargeable energy solution. Through real-time data monitoring, it optimizes the collection schedule, reduces health risks associated with overflowing bins, and promotes a cleaner campus environment.

Keywords: Smart Waste Bin, IoT, Waste Alert System, TEMAC, SHRUMAC

I. INTRODUCTION

Since life itself and the creation of man, waste has been part of human existence. It is defined as any by-product, material, or substance that is discarded because it is no longer useful or needed; these include industrial by-products, sewage, and garbage; hence, most human daily activities generate waste [1]. Some of these actions stem from extracurricular activities such as eating and drinking. These generated wastes must be properly managed and disposed of to protect human health and the environment at large. Waste exists in many forms, and its characterization can be expressed in various ways. They are often located in different environments and landscapes [2]. To reduce the generation and spread of these human-generated wastes, various equipment, methods, and tools were used, taking environmental differences into account. This equipment includes garbage bins, brooms, parkers, garbage trucks, and disposable bags. Individuals often manage this equipment. Waste is easily generated in congested areas such as schools, markets, hospitals, and worship centers. As individuals consume different products and services, ranging from biodegradable to non-biodegradable items such as food, plastic, and paper. These materials and products are usually discarded in waste bins and sometimes on the ground. When these items accumulate over time, they pile up; this waste becomes unpleasant to the eyes and nose and can even pose health risks. To reduce environmental pollution and improve individual health, proper trash disposal practices are essential for ensuring overall environmental well-being [3]. As stated by [4].

Insufficient waste management infrastructure, limited awareness of proper waste disposal practices, and inadequate institutional policies may contribute to suboptimal waste management outcomes. Though environmental cleanliness and sanitation are everyone's responsibility, specialized individuals, known as cleaners, are employed to collect trash and waste regularly. These individuals are responsible for locating, collecting, and removing waste materials from strategic locations to keep the environment clean at all times. They set up waste bins throughout the environment to collect waste and mark certain locations for easy retrieval. It takes a significant number of waste bins and a variety of emptying schedules to handle the public waste effectively. However, certain factors, such as unidentified, piled-up waste bins and the location of individual bins, could hinder the effective handling and removal of these materials. Due to regular patronage or population in a location, some waste bins fill up much faster than they are emptied, leading to them becoming full before the next collection date or time, which poses hygiene risks to the public [5][6][7]. Given these challenges, there is a pressing need for a better approach to sanitary and waste management systems, and for their adoption or inclusion to achieve a more efficient, healthier, and more hygienic society. The smart waste management system will not only shed light on the challenges faced by cleaners. Still, it will also pave the way for targeted interventions to enhance the sustainability and effectiveness of sanitation efforts. The introduction and use of a real-time sanitary and waste alert system to help address most of the challenges encountered in gathering, isolating, managing, and disposing of waste are the focal points of this research. The Internet of Things (IoT), that is, the interconnection of devices through the internet, presents a revolutionary approach to enhancing waste management practices. IoT integrates sensors, communication networks, and data analytics to create intelligent systems that monitor, analyze, and optimize various operations remotely and in real time [8]. About waste management, IoT technology helps create smart trash bins that monitor fill levels, detect hazardous gas emissions, and provide real-time data to waste management authorities via mobile alert systems, enabling timely interventions. By equipping waste bins with sensors to monitor fill levels and hazardous gas emissions, IoT can optimize waste collection, prevent environmental hazards, and enable real-time data-driven decision-making for municipalities and waste management authorities across campuses of Nigerian higher institutions. Higher institutions of learning generate substantial amounts of municipal solid waste daily due to academic activities, commercial operations, student residences, cafeterias, and administrative offices. The increasing population of students and staff has further intensified the challenge of managing waste effectively within university campuses. In many developing countries, waste bins often overflow before they are emptied because cleaners rely on routine inspection rather than real-time information about bin status. This situation contributes to environmental pollution, unpleasant odors, pest infestation, and increased public health risks. According to [9]. Sustainable waste management requires efficient collection systems supported by appropriate technologies to minimize environmental degradation. Similarly, [10]. observed that ineffective monitoring of waste accumulation significantly reduces the efficiency of municipal solid waste management systems and increases operational costs. The emergence of the Internet of Things (IoT) has created new opportunities for improving waste management through intelligent monitoring and automation. IoT-based waste management systems integrate sensors, embedded controllers, communication technologies, and mobile applications to continuously monitor waste levels and notify responsible personnel when intervention is required. Such systems eliminate unnecessary manual inspection while ensuring that waste is collected before bins overflow. A comprehensive review by [11]. concluded that smart waste bins equipped with ultrasonic sensors and wireless communication significantly improve waste collection efficiency, optimize resource utilization, and promote environmental sustainability. Likewise, [12]. demonstrated that IoT-enabled smart waste monitoring systems provide real-time information on waste-bin status and support cleaner urban environments through automated monitoring and timely notification. Despite these technological advancements, the adoption of intelligent waste management systems remains very limited in many Nigerian tertiary institutions, where conventional waste bins are still predominantly used. Most campus waste management practices rely on periodic inspections by cleaners, leading to delayed waste collection, overflowing bins, and poor sanitation. Furthermore, conventional waste bins require users to touch the lid during disposal, thereby increasing the risk of transmitting disease-causing microorganisms

via contaminated surfaces. Reference [1] emphasized that modern waste management systems should incorporate technological innovations that improve operational efficiency and environmental hygiene. Similarly, [13]. argued that sustainable waste management requires the integration of smart technologies capable of supporting timely decision-making and improving sanitation outcomes. It is against this backdrop that the SHRUMAC (Smart Hygiene and Refuse Monitoring and Alert Controller) system was developed as an interventionist mechanism to address the shortcomings of existing campus waste management practices. SHRUMAC integrates an ultrasonic sensor, an ESP32 microcontroller, Bluetooth communication, a servo motor, and a companion mobile application, TEMAC (Technology-Enabled Monitoring and Alert Controller), into a single intelligent waste-monitoring platform. The system continuously monitors waste bin fill levels, automatically opens the lid when a user approaches, and sends real-time notifications to cleaners when waste reaches a predetermined threshold. This approach reduces unnecessary manual inspections, prevents waste overflow, promotes hygienic waste disposal, and enhances the overall efficiency of campus sanitation management. The effectiveness of similar IoT-based waste-monitoring architectures has been demonstrated in [14]. who reported improved responsiveness of waste collectors through mobile notifications, while [12]. found that integrating sensing, automation, and wireless communication significantly enhances the performance of smart waste management systems.

1. Statement of the Problem

Effective waste management remains a major environmental and public health challenge in many higher educational institutions, particularly in developing countries. Although waste bins are strategically positioned across university campuses to promote proper waste disposal, they are typically monitored through routine manual inspections rather than real-time systems. Consequently, waste bins frequently become overfilled before they are emptied, resulting in unpleasant odor, littering, pest infestation, environmental pollution, and increased risks of disease transmission. Reference [9] observed that inefficient waste collection practices and inadequate monitoring systems are among the major factors affecting sustainable waste management in developing countries. Similarly, [10]. reported that the absence of timely information on waste accumulation contributes significantly to poor waste collection efficiency and increased operational costs. Another concern associated with conventional waste bins is their dependence on manual operation. Users are often required to touch the waste bin lid during disposal, creating additional opportunities for the transmission of bacteria, viruses, and other harmful microorganisms via contaminated surfaces. Furthermore, sanitation personnel lack reliable means to determine when waste bins are at capacity, making waste collection largely dependent on fixed schedules or physical inspection. This approach is time-consuming, labor-intensive, and often results in delayed waste evacuation and overflowing bins. Recent studies on smart waste management have shown that integrating sensors, embedded controllers, wireless communication, and mobile applications can provide real-time monitoring of waste bins, improve collection efficiency, and promote cleaner environments [12][14]. Despite the increasing adoption of Internet of Things (IoT)-based waste management technologies worldwide, their use in Nigerian tertiary institutions remains limited. Most campuses still rely on conventional waste bins that lack automated monitoring, real-time notification, and touch-free operation. This technological gap reduces the effectiveness of campus sanitation programs and limits cleaners' ability to respond promptly to waste management needs. It is therefore necessary to develop an intelligent waste management system that can automatically monitor waste levels, detect user presence, enable contactless lid operation, and transmit real-time alerts to sanitation personnel. The development of SHRUMAC (Smart Hygiene and Refuse Monitoring and Alert Controller), integrated with the TEMAC (Technology-Enabled Monitoring and Alert Controller) mobile application, is intended to address these shortcomings by providing an affordable, efficient, and smart waste monitoring solution that supports timely waste collection, improves environmental hygiene, and enhances sanitation management within educational institutions.

2. Aim and Objectives

This research aimed to develop a novel, real-time IoT-based sanitary and waste alert system to enhance cleanliness and efficiency on university campuses.

This was achieved by:

Developing SHRUMAC, a hardware system (using an Arduino microcontroller) that will process data from all connected sensors, manage wireless communication, and send alerts to the mobile phone application and user devices, ensuring continuous monitoring and consistent communication. Building TEMAC, a Mobile application using MIT App Inventor that will be connected to the system's hardware remotely. Implement an alert mechanism using an alarm buzzer and a mobile application notification to enable the system to send real-time alerts to designated personnel via the mobile application interface, ensuring timely response to issues such as full waste bins. Adopting a rechargeable energy solution using batteries to run the system to ensure an uninterrupted power supply for both indoor and outdoor installations. Testing. Evaluation and deployment of the system.

II. LITERATURE REVIEW

This section provides a comprehensive overview of the evolution and current state of waste management systems, detailing the transition from early, often hazardous disposal methods to modern technology-driven practices. It explores the historical context, highlighting the progression from rudimentary techniques in ancient Rome and medieval Europe to systematic approaches prompted by urbanization and industrialization. Technological advancements, including automated collection systems, AI-driven sorting, and waste-to-energy technologies, are examined for their role in improving efficiency and sustainability. Current trends emphasize circular economy models, extended producer responsibility, and smart waste management solutions. It concludes by summarising related literature on smart waste management systems, environmental impact, public health, and technological innovations, setting the stage for further research and developments.

Evolution of Waste Management Systems (Early Waste Management Practices)

Early waste management practices varied greatly across different cultures and regions, influenced by local conditions and available resources. In many ancient societies, waste was often disposed of haphazardly, leading to significant health and environmental issues. For instance, in ancient Rome, despite a relatively sophisticated system of aqueducts and sewers, waste management practices were still rudimentary and largely ineffective at preventing the spread of disease [10]. In medieval Europe, waste disposal methods were even less advanced. Garbage and human waste were frequently discarded into the streets, rivers, and other public places, contributing to the spread of plagues and other infectious diseases [13]. It wasn't until the 19th century that more systematic approaches to waste management began to emerge, largely in response to growing urbanization and public health concerns. In contrast, some early civilizations, such as those in ancient China and Japan, practiced more organized waste management. These societies employed methods such as composting and recycling organic waste for agricultural purposes, demonstrating an early understanding of the potential value of waste [15]. These practices were driven by necessity, as densely populated areas required more efficient waste management to prevent disease outbreaks and sustain agricultural productivity. The Industrial Revolution marked a significant turning point in waste management practices. The massive increase in waste production due to industrial activities necessitated the development of new disposal methods, including the establishment of landfills and incineration plants [9]. These methods, while more organized, often led to environmental degradation and

pollution, highlighting the need for more sustainable waste management solutions. In summary, early waste management practices were largely reactive and localized, evolving from rudimentary and often hazardous methods to more structured approaches in response to urbanization and industrialization. The varied practices across different cultures underscore the importance of context-specific solutions in waste management [16][17].

1. Technological Advancements

Technological advancements in waste management have significantly transformed how societies handle waste, making processes more efficient, sustainable, and environmentally friendly. Over the past few decades, innovations in waste management technology have addressed various challenges, including waste collection, sorting, recycling, and disposal. One of the most notable advancements is the development of automated waste collection systems. These systems use sensors and smart bins to optimize waste collection routes, reducing fuel consumption and operational costs [18]. For instance, smart bins equipped with fill-level sensors notify waste collection services when they need emptying, ensuring timely and efficient waste collection. In the realm of recycling, chemical recycling technologies have emerged as a game-changer. Unlike traditional mechanical recycling, which can degrade material quality, chemical recycling breaks down plastics into their molecular components, enabling the production of new, high-quality materials. This process not only improves the recyclability of complex plastic products but also helps in managing plastic waste more sustainably [19]. Waste-to-Energy (WtE) technologies have also made significant strides. Modern incineration plants are equipped with advanced emission control systems that minimize environmental pollution while generating electricity and heat from waste. Additionally, aerobic digestion technologies convert organic waste into biogas, a renewable energy source that reduces reliance on fossil fuels and lowers greenhouse gas emissions [20]. Another significant advancement is the use of digital platforms and big data analytics in waste management. These technologies enable municipalities and waste management companies to monitor and analyze waste generation patterns, optimize resource allocation, and improve decision-making processes. Predictive analytics, for instance, can forecast waste-generation trends, helping plan for future waste-management needs [21]. In summary, technological advancements in waste management have revolutionized how waste is collected, sorted, recycled, and disposed of. Innovations such as automated collection systems, advanced sorting technologies, chemical recycling, waste-to-energy processes, and digital platforms have made waste management more efficient, sustainable, and environmentally friendly [22].

2. Current Trends

Current trends in waste management reflect the growing emphasis on sustainability, circular economy principles, and the integration of advanced technologies to manage waste more efficiently. These trends highlight the shift from traditional waste disposal methods to more innovative and environmentally friendly practices. One prominent trend is the increasing adoption of circular economy models. Unlike the linear "take-make-dispose" approach, the circular economy aims to keep resources in use for as long as possible, extract the maximum value from them, and recover and regenerate products and materials at the end of their lifecycle. This approach reduces waste generation and promotes resource efficiency [23]. Many countries and businesses are implementing circular economy strategies to minimize waste and enhance sustainability. The rise of extended producer responsibility (EPR) is another significant trend. EPR policies hold manufacturers accountable for the entire lifecycle of their products, including end-of-life disposal. This encourages producers to design products that are easier to recycle, reuse, or dispose of safely. EPR programs have been successful in improving recycling rates and reducing environmental impacts across sectors such as electronics, packaging, and automotive [24]. Technological advancements continue to drive innovation in waste management. The use of big data and Internet of Things (IoT) technologies enables real-time monitoring and optimization of waste collection and processing

systems. Smart waste management solutions, such as sensor-equipped bins and intelligent routing systems, help reduce operational costs and environmental footprints by optimizing waste collection schedules and routes [25]. Waste-to-energy (WtE) technologies remain a crucial component of modern waste management strategies. These technologies convert non-recyclable waste materials into usable forms of energy, such as electricity, heat, or fuel. Advanced WtE plants incorporate cutting-edge emission control systems to minimize environmental impacts, making them a viable solution for managing residual waste while generating renewable energy [26]. Another emerging trend is the focus on bio-waste management. Organic waste, which includes food waste, garden waste, and agricultural residues, constitutes a significant portion of municipal solid waste. Innovative bio-waste treatment technologies, such as anaerobic digestion and composting, are being increasingly adopted to convert organic waste into valuable products like biogas and compost. These processes not only reduce landfill usage but also contribute to renewable energy production and soil health [27]. Plastic waste management has garnered substantial attention in recent years, driven by the global plastic pollution crisis. Public awareness campaigns and international agreements, such as the Basel Convention's amendments on plastic waste, are also playing a crucial role in tackling plastic pollution [28]. Therefore, community engagement and public participation are becoming increasingly important in waste management. In conclusion, current trends in waste management emphasize sustainability, technological innovation, and community involvement. The shift towards circular economy models, extended producer responsibility, smart waste management solutions, waste-to-energy technologies, bio-waste management, plastic waste reduction, and public engagement are shaping the future of waste management, making it more efficient, sustainable, and environmentally friendly [23][25][28].

3. The SHRUMAC System

The SHRUMAC system is designed to enhance waste management through automation and real-time monitoring. It integrates several components to ensure efficient operation and user convenience. Ultrasonic Waste Level Measurement: An ultrasonic sensor continuously monitors the trash can's fill level. This sensor emits ultrasonic waves, which reflect off the waste material back to the sensor, allowing it to measure the distance from the sensor to the top of the waste. By analyzing these reflections, the system can determine how full the trash can is. Bluetooth Communication: When the trash can reaches a certain fill threshold, the ultrasonic sensor triggers a notification. This notification is transmitted via Bluetooth to the TEMAC, a mobile application. The TEMAC App, developed using MIT App Inventor, receives this alert and notifies the user that the waste needs to be disposed of or the bin needs to be managed. This feature ensures that users are promptly updated about the waste status without needing to check the bin manually. Automatic Lid Operation: To enhance hygiene and convenience, the trash can is equipped with a servo motor controlled by an ultrasonic sensor. When a user approaches the bin, the sensor detects their presence and automatically opens the lid. This hands-free operation reduces the need for physical contact with the bin, thereby promoting better hygiene and a more user-friendly experience. System Integration: The integration of these components, the ultrasonic sensor, Bluetooth communication, mobile app, and servo motor, creates a cohesive system that not only alerts users about waste levels but also provides an automated solution for lid operation. This synergy helps maintain a cleaner environment and improve the overall efficiency of waste management.

Overall, the proposed system combines advanced sensing technology with automated mechanisms to provide an innovative solution for the modern waste management challenge.

4. Advantages of the SHRUMAC System

Improved Operational Efficiency: Implementing a real-time sanitary and waste alert system offers numerous benefits. One primary benefit is improved operational efficiency. The system’s real-time monitoring ensures

timely waste collection, reducing the risk of overflow and maintaining a clean environment. This proactive approach reduces emergency responses and ad hoc collections, optimizing resource allocation and lowering operational costs. Reduced Environmental Impact: The system promotes environmental sustainability by ensuring timely and efficient waste collection. By optimizing collection routes and schedules based on real-time data, fuel consumption and associated emissions are minimized, thereby reducing the carbon footprint of waste collection operations. Additionally, accurate data on waste types and quantities facilitates targeted recycling programs, reducing the amount of waste sent to landfills and promoting recycling efforts. This targeted approach aligns with global sustainability goals and supports environmental protection initiatives. Enhanced Public Health and Safety: A real-time sanitary and waste alert system significantly improves public health and safety. Timely waste collection prevents the accumulation of waste, which can harbor pathogens and attract pests, thereby posing health risks to the community. The system can also detect hazardous waste, enabling immediate action to mitigate potential health hazards. This proactive approach ensures a safer and healthier environment for the community, reducing the incidence of waste-related health issues. Data-Driven Decision Making: The system provides valuable data on waste-generation patterns to inform data-driven decision-making. Authorities can analyze trends and patterns to implement effective waste management strategies. For instance, identifying peak waste-generation times can help schedule additional collections during those periods. Data analytics can also help in identifying areas with higher waste generation, enabling targeted interventions and resource allocation. Cost Savings: Optimizing waste collection routes and schedules yields significant cost savings. Reduced fuel consumption and operational costs, combined with efficient resource use, contribute to overall cost reduction. The long-term savings in operational expenses can offset the initial investment in the system. Additionally, the system reduces the need for emergency collections and unplanned maintenance, further contributing to cost savings. Community Engagement and Awareness: Implementing a real-time sanitary and waste alert system can enhance community engagement and awareness. By providing real-time data and alerts, the system encourages community participation in waste management efforts. Awareness campaigns and educational programs can be designed based on the collected data to promote responsible waste disposal practices and environmental consciousness. Engaging the community fosters a sense of ownership and responsibility towards maintaining a clean and healthy environment.

5. Architecture of the SHRUMAC System

The proposed system, named SHRUMAC (Smart Hygiene and Refuse Monitoring and Alert Controller), is an intelligent waste-monitoring device designed to improve sanitation management through real-time waste detection and automated notifications. SHRUMAC consists of several integrated hardware and software components that work together to monitor the waste level in a trash bin, automatically operate the lid, and communicate the bin's status to a companion mobile application known as TEMAC (Technology-Enabled Monitoring and Alert Controller). The sensing unit forms the first component of the system. It employs an ultrasonic sensor mounted at the top of the waste bin to continuously measure the distance between the sensor and the waste surface. As refuse accumulates in the bin, the measured distance decreases. These distance readings are transmitted to the microcontroller for further processing. The same ultrasonic sensor detects a user approaching the bin, enabling touch-free operation of the lid. The microcontroller serves as the central processing unit of SHRUMAC. An ESP32 microcontroller is recommended because it provides sufficient processing capability and has an in-built Bluetooth communication module. The microcontroller receives distance measurements from the ultrasonic sensor, processes the data, and calculates the current fill level of the waste bin. It then compares the measured fill level against predefined thresholds to determine whether the bin is empty, partially filled, nearly full, or full. Once the fill level reaches the specified threshold, the microcontroller generates an alert message for transmission to the mobile application. Communication between SHRUMAC and the user is via Bluetooth. The Bluetooth module, integrated into the ESP32 microcontroller, wirelessly transmits processed data from the waste monitoring system to the TEMAC mobile application. This communication

enables cleaners or sanitation personnel to receive immediate notifications when the waste bin approaches capacity, ensuring prompt waste evacuation and preventing overflow. The TEMAC mobile application serves as the user interface of the proposed system. Developed using MIT App Inventor, the application receives real-time notifications from SHRUMAC and displays important information such as the current fill level of the waste bin and its operational status. The application also allows sanitation personnel to monitor the condition of the waste bin and respond promptly whenever cleaning or waste disposal is required. This eliminates the need for manual inspection of waste bins and enhances the efficiency of sanitation operations. To improve hygiene and convenience, the proposed system incorporates an automatic lid operation mechanism. Whenever the ultrasonic sensor detects a user within a predetermined distance, the microcontroller automatically activates a servo motor to open the waste bin lid. After the waste has been disposed of, the microcontroller instructs the servo motor to close the lid after a short delay. This touch-free operation minimizes human contact with the waste bin and improves environmental hygiene. The entire SHRUMAC system is powered by a regulated power supply, which may consist of a rechargeable battery or an external power adapter, depending on the deployment environment. The power supply provides stable electrical power to the ultrasonic sensor, microcontroller, Bluetooth module, servo motor, and other electronic components, ensuring reliable, uninterrupted operation of the system.

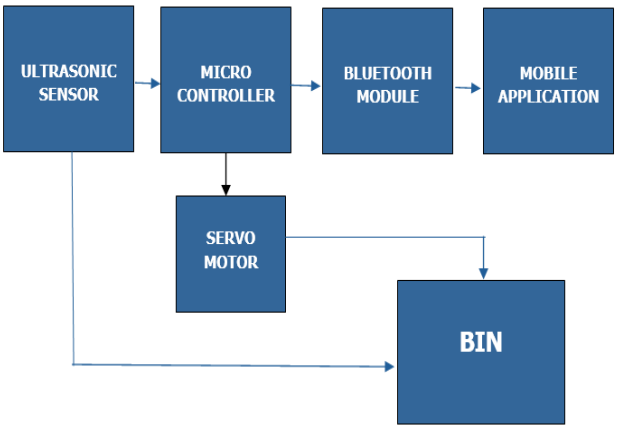


Figure 1:Architecture of the SHRUMAC System

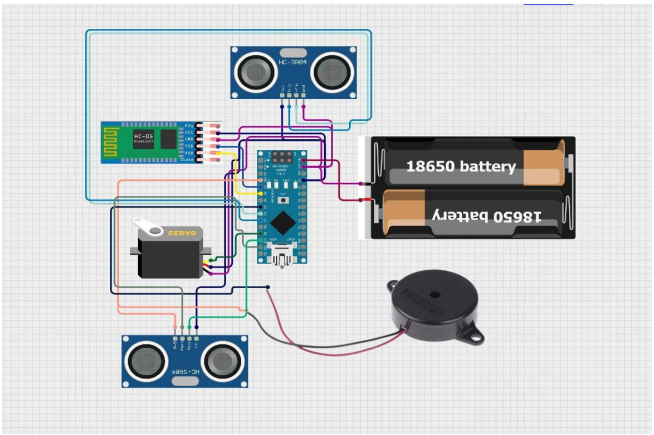


Figure 2. Circuit Diagram of the SHRUMAC system showing circuit connections for sensors, microcontroller, and other components

III. RESULTS





Figure 3. Waste bin used in the development of Figure 4. A servomotor is used for the automation of the *SHRUMAC system for* the SHRUMAC lid operation



Figure 5. Ultrasonic sensor used in the measurement of waste levels in the SHRUMAC system



Figure 6. Buzzer when the SHRUMAC system is full

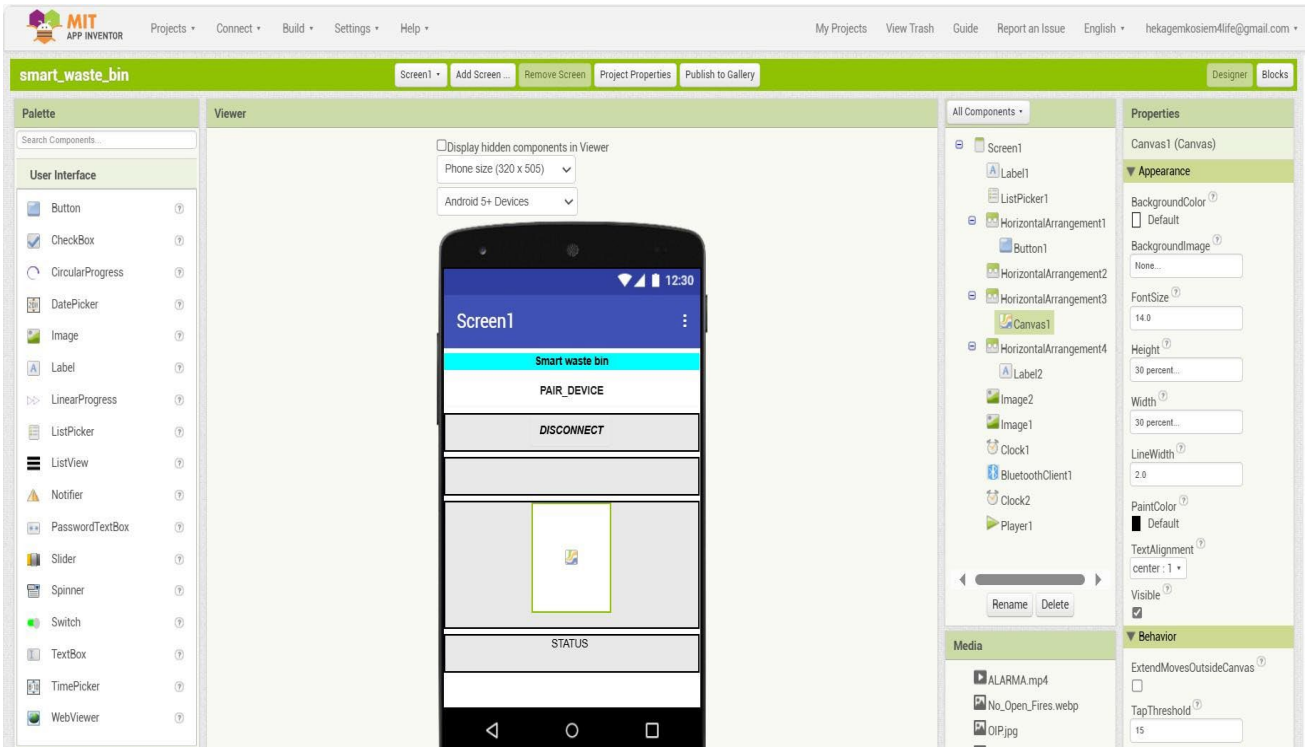


Figure 7. Frontend interface of the mobile App (TEMAC)



Figure 8. Hardware assembly displaying the assembled hardware components of the SHRUMAC system.



Figure 9. Testing the SHRUMAC Device

1. Summary of Findings

The SHRUMAC system successfully met its design objectives, demonstrating effective performance in waste detection, lid control, and user notification. The system's integration of hardware and software components allowed it to operate efficiently and address user needs effectively. The accuracy of the ultrasonic sensors, the quick response of the servo motor, and reliable battery performance highlight the system's potential for practical application in waste management.

2. Discussion of Findings

The findings of this study indicate that the SHRUMAC (Smart Hygiene and Refuse Monitoring and Alert Controller) successfully achieved its intended objectives by providing an effective smart waste monitoring solution. The system accurately detected the waste level using ultrasonic sensing technology, automatically controlled the waste-bin lid via a servo motor, and sent timely notifications to users through the TEMAC (Technology-Enabled Monitoring and Alert Controller) mobile application. The seamless integration of these hardware and software components demonstrates that embedded systems can be effectively employed to improve sanitation management and promote environmental cleanliness. The study revealed that the ultrasonic sensor consistently measured the waste bin's fill level with satisfactory accuracy, enabling the system to distinguish among different waste levels and trigger alerts when predetermined thresholds were reached. This finding agrees with the work of [14]. who reported that ultrasonic sensors integrated with an ESP32 microcontroller provided reliable monitoring of waste volume and generated real-time notifications to waste collectors, thereby reducing delays in waste evacuation. Similarly, the systematic review conducted by [11]. identified ultrasonic sensors as the most widely adopted sensing technology in smart garbage bins because of their accuracy, affordability, and ease of integration with embedded controllers for fill-level monitoring. The automatic lid operation of SHRUMAC also proved effective. The servo motor responded promptly whenever the ultrasonic sensor detected a user, allowing the waste-bin lid to open and close without physical contact. This touch-free mechanism improves user convenience while reducing the risk of transmitting germs through

contaminated surfaces. The finding supports the work of [12]. who developed an IoT-based smart waste monitoring system that combined an ultrasonic sensor and a servo motor to provide automatic lid operation and real-time waste monitoring. Likewise, reported that integrating ultrasonic sensors with servo motors enhanced hygiene by enabling contactless waste disposal and monitoring waste levels for efficient collection. Another important finding of the study is the effectiveness of the communication mechanism between SHRUMAC and the TEMAC mobile application. The Bluetooth-based notification system successfully informed users whenever the waste bin reached the predefined fill level, thereby allowing cleaners to respond promptly before overflow occurred. This finding is consistent with previous studies on smart waste management systems, which have shown that real-time communication between smart bins and monitoring applications significantly improves waste collection efficiency and reduces unnecessary inspection visits. According to [11]. communication technologies such as Bluetooth, Wi-Fi, and Zigbee play a crucial role in enabling smart garbage bins to exchange real-time information with users and waste management personnel. Similarly, [14]. demonstrated that mobile notifications greatly improved the responsiveness of waste collectors by providing immediate information on bin status. The findings further showed that the battery-powered SHRUMAC prototype maintained stable operation throughout the testing period, with no noticeable interruptions. Reliable power performance is essential for embedded systems intended for continuous environmental monitoring, particularly where uninterrupted sensing and communication are required. The system's overall stability demonstrates that SHRUMAC has the functional reliability needed for deployment in schools, offices, hospitals, markets, residential estates, and other public environments where efficient waste management is essential. Summarily, the findings confirm that SHRUMAC is a practical and efficient smart waste management system capable of addressing common challenges associated with conventional waste bins, including delayed waste collection, overflowing bins, and unhygienic manual lid operation. By combining accurate sensing, automated control, wireless communication, and a user-friendly mobile application, the proposed system provides a cost-effective solution that supports cleaner environments and aligns with the growing adoption of Internet of Things (IoT)-based technologies for sustainable waste management.

3. Recommendations

Based on the findings of this research, the following recommendation is made: Researchers should conduct further studies to refine and improve sensor accuracy, particularly under varying environmental conditions. Future researchers should focus on Battery Life optimisation by exploring alternative power sources. Future researchers may consider expanding the integrated features, such as data analytics and advanced user interfaces, to further enhance functionality and user experience.

REFERENCES

[1]. Brunner, P. H., & Rechberger, H. (2015). Energy waste—Key element for sustainable waste management. Waste Management, 37, 3–12. https://doi.org/10.1016/j.wasman.2014 .02.003

[2]. Amasuomo, E., & Baird, J. (2016). Waste characterization: A useful tool for the development of waste management strategies. Waste Management, 47, 22–28.

[3]. Smith, R. J. (2021). Importance of proper waste disposal practices for maintaining public health. Public Health Review, 29(1), 78–85.

[4]. Ugwu, C. O., Ozoegwu, C. G., & Ozor, P. A. (2020). Solid waste quantification and characterization in the University of Nigeria, Nsukka campus, and recommendations for sustainable management. Heliyon, 6(6), e04255. https://doi.org/10.1016/j.heliyon. 2020.e04255

[5]. MohdYusof, N., Abdul Ghafar, N., & Sulaiman, Z. (2017). Factors affecting the effectiveness of waste management systems: A case study of a university campus. International Journal of Waste Resources,

7(3), 215–225.

[6]. Zeb, R., Awan, U., & Abbasi, S. (2019). The impact of waste management practices on environmental sustainability in educational institutions. Sustainable Environment Research, 29, Article 12. https://doi.org/10.1186/s42834-019-0012-0

[7]. Harith, Z. T., Subramaniam, P., & Lai, K. W. (2020). Challenges in implementing effective waste management systems in university campuses. Journal of Environmental Management and Sustainability, 14(4), 45–59.

[8]. Abidemi, O, Olamide, F., Adeyinka, A., & Alashiri, O. (2022). Design and development of a smart waste bin for effective waste collection. In Future Trends in Networks and Computing Technologies (pp. 363–372). Springer. https://doi.org/10.1007/978-981-19-5037-7_26

[9]. Wilson, D. C., Velis, C. A., & Rodic, L. (2013). Integrated sustainable waste management in developing countries. Proceedings of the Institution of Civil Engineers – Waste and Resource Management, 166(2), 52–68. https://doi.org/10.1680/warm.12.00005

[10]. Sosunova, I., and Porras, J. (2022). IoT-Enabled Smart Waste Management Systems for Smart Cities: A Systematic Review, in IEEE Access, vol. 10, pp. 73326-73363, 2022, doi: 10.1109/ACCESS.2022.3188308.

[11]. Aurko, N. S., Masum, S. M., & Hossain, M. S. (2024). An IoT-based smart waste monitoring system envisions a sustainable smart city. In the 2024 IEEE Conference on Computing Applications and Systems (COMPAS). https://doi.org/10.1109/COMPAS60761. 2024.10796021

[12]. David, V. E., Jr., John, Y., & Hussain, S. (2020). Rethinking sustainability: A review of Liberia's municipal solid waste management systems, status, and challenges. Journal of Material Cycles and Waste Management, 22(4), 1299–1317.

[13]. Widodo, & Rifky, A. N. (2021). IoT-based smart trash monitoring using the Blynk application. BEST: Journal of Applied Electrical, Science and Technology, 3(2), 25–29. https://doi.org/10.36456/best.v3i2.4266

[14]. Taherzadeh, M.; Rajendran, K. (2015). Factors Affecting the Development of Waste Management: Experience from Different Cultures. In Waste Management and Sustainable Consumption: Reflections on Waste; Estrom, K.M., Ed.; Routledge: Abingdon, UK.

[15]. Habib, M. S., Sarkar, B., Tayyab, M., Saleem, M. W., Hussain, A., Ullah, M., Omair, M., & Iqbal, M. W. (2019). Large-scale disaster waste management under an uncertain environment. Journal of Cleaner Production, 212, 200–222. https://doi.org/10.1016/j.jclepro.2018.11.267

[16]. Hoang, N.H.; Fogarassy, C. (2020). Sustainability evaluation of municipal solid waste management system for Hanoi (Vietnam) - why to choose the "waste-to-energy" concept—sustainability 12, 1085.

[17]. Krook, J., Carlsson, A., & Malmström, M. E. (2018). Urban mining and landfill mining: Emerging global perspectives and approaches to waste management. Journal of Cleaner Production, 197, 1365–367. https://doi.org/10.1016/j.jclepro.2018.06.050

[18]. Hahladakis, J.N.; Iacovidou, E. (2018). Closing the loop on plastic packaging materials: What is quality and how does it affect their circularity? Sci. Total Environ; 630, 1394–1400.

[19]. Tchobanoglous, G.; Kreith, F. (2002). Handbook of Solid Waste Management; McGraw-Hill: New York, NY, USA.

[20]. Zaman, A.U., & Lehmann, S. (2013). The zero waste index: A performance measurement tool for waste management systems in a ‘zero waste city’. Journal of Cleaner Production, 50, 123-132.

[21]. Zhou, G., Lim, M.K., Tseng, M.L., & Bui, T.D. (2014). Promoting sustainability in the construction industry supply chain: Cases in developing countries. J. Clean. Prod., 142, 1518–1536.

[22]. Geissdoerfer, M., Savaget, P., Bocken, N.M.P., & Hultink, E.J. (2017). The Circular Economy – A new sustainability paradigm? Journal of Cleaner Production, 143, 757-768.

[23]. OECD. (2016). Extended Producer Responsibility: Updated Guidance for Efficient Waste Management. OECD Publishing, Paris.

[24]. Pitt, M., Puckering, D., & Chen, J. (2016). Smart waste management: A smartphone application as a means of reducing food waste. In Proceedings of the 11th International Conference on Waste Management and Technology.
[25]. Arena, U. (2012). Process and technological aspects of municipal solid waste gasification. A review.
[26]. Kothari, R., Tyagi, V.V., & Pathak, A. (2010). Waste-to-energy: A way from renewable energy sources to sustainable development. Renewable and Sustainable Energy Reviews, 14(9), 3164-3170.
[27]. Geyer, R., Jambeck, J.R., & Law, K.L. (2017). Production, use, and fate of all plastics ever made. Science Advances, 3(7), e1700782.
[28]. Telusoori, R. B., Harika, K., Gnanitha, S., Dileep, B., & Vivek, P. (2026). Intelligent ESP32-based IoT waste management system for efficient waste collection solutions. International Journal of Engineering Research and Science & Technology, 22(2), 815–821.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

From Textual Variation to Functional Diversity in Beginner C Programming Tasks Generated by Ollama Cloud-Based Large Language Models

Ladislav Végh
Department of Informatics
J. Selye University
Komárno, Slovakia
veghl@ujs.sk
ORCID:0000-0001-9242-4983

Dávid Demeter
Department of Informatics
J. Selye University
Komárno, Slovakia
127616@student.ujs.sk
ORCID:0009-0001-2016-6708

Ondrej Takáč
Department of Informatics
J. Selye University
Komárno, Slovakia
takaco@ujs.sk
ORCID:0000-0001-8836-0227

Krisztina Czakóová
Department of Informatics
J. Selye University
Komárno, Slovakia
czakoovak@ujs.sk
ORCID:0009-0000-3096-9427

Abstract
Large language models can generate programming exercises, but textually different outputs often do not represent genuinely different tasks. This distinction is important because a large set of differently worded exercises may still train the same computational operation, giving instructors and learners the impression of broad task coverage while providing limited functional variety. This study evaluated textual diversity, functional diversity, and output reliability in LLM-generated beginner-level C programming tasks. Six Ollama cloud-based models were tested across four prompt categories, four temperature settings, and ten repeated generations per condition, yielding 960 outputs. Textual diversity was calculated as the proportion of distinct task descriptions, and functional diversity as the proportion of manually verified functional groups based on the required programming operation. Across models, textual diversity ranged from 53% to 88%, whereas functional diversity ranged from 10% to 32%. The functional-to-textual diversity ratio remained below 0.5 for all models, indicating that many differently worded tasks required the same underlying operation. Temperature increased textual diversity more strongly than functional diversity: DeepSeek V4, Qwen3, and Kimi K2.6 reached 100% textual diversity at higher temperatures, while functional diversity remained lower. Prompt category had a stronger effect on functional diversity than temperature: modification tasks produced the highest functional diversity (40%), whereas search tasks produced the lowest (10%). Output reliability was high, with five models producing valid JSON in all cases and one model reaching 99.4%. However, markdown code block usage varied substantially, reaching 84.4% for Gemma 4. The findings suggest that LLM-based programming task generation should be evaluated not only by textual distinctiveness but also by human-supervised semantic validation, functional coverage, and output validity.

Keywords: large language models; programming education; task generation; textual diversity; functional diversity; prompt engineering; output reliability

I. INTRODUCTION

Large language models (LLMs) are increasingly used in computer science education for code explanation, feedback generation, assessment support, and for the automatic generation of programming exercises [1], [2]. In

introductory programming courses, exercise generation is especially relevant because students need repeated practice with similar concepts, while instructors need many short, solvable, and appropriately constrained tasks. However, the educational value of automatically generated tasks does not depend only on whether the outputs are linguistically different. A set of tasks may contain many different textual formulations while still requiring students to solve the same programming problem. This creates a practical problem for question generation in educational settings. If an LLM generates many textually distinct tasks that all require the same operation, such as counting selected elements, finding a minimum value, or modifying array elements based on a condition, the resulting task set may appear diverse yet offer only limited conceptual coverage. For instructors, this may lead to an overestimation of the variety of automatically generated exercises. For students, it may result in repeated practice of the same type of functional task rather than exposure to a broader range of beginner-level programming operations. Therefore, low functional diversity relative to textual diversity is not only a measurement issue but also a pedagogical concern. Previous work has emphasized the need for systematic evaluation, quality assurance, and controlled generation conditions [3], [4], [5], [6], [7]. In parallel, studies of LLM diversity and creativity have examined how sampling parameters, prompting strategies, and model behavior influence output variation [8], [9], [10]. However, in programming task generation, the relationship between textual diversity and functional diversity remains insufficiently examined. The present study addressed this issue by evaluating beginner-level C programming tasks generated by six Ollama cloud-based LLMs under controlled prompting conditions. The prompts targeted four task categories: aggregation, search, modification, and two-array tasks. These categories were selected because they represent common beginner-level array-processing patterns in introductory programming courses. Aggregation tasks require students to process all elements of an array and reduce them to a single result, such as a sum, count, minimum, maximum, or average. Search tasks require locating or identifying elements that satisfy a given condition. Modification tasks focus on transforming array elements according to a rule, while two-array tasks require coordinated processing of two input arrays. Together, these categories cover frequently taught functional patterns while keeping the required programming constructs limited to loops, conditionals, and arrays. This made it possible to compare textual and functional diversity across pedagogically relevant task types while maintaining controlled complexity. The study distinguished between textual diversity, calculated from the number of distinct task descriptions, and functional diversity, calculated from the number of semantically validated functional groups representing the underlying programming task types. In addition, output reliability was assessed using JSON validity, markdown code block usage, and truncation rate, as these factors affect the practical use of LLM outputs in automated educational pipelines. The following research questions guided the study:

- How does sampling temperature influence textual diversity in beginner programming task generation?
- How does sampling temperature influence functional diversity in beginner programming task generation?
- To what extent does increased textual diversity correspond to genuinely new functional task types?
- How do model choice and prompt category affect textual and functional diversity?

The main contribution of the paper is the explicit comparison between textual and functional diversity in LLM-generated programming tasks. Rather than treating variation in wording as sufficient evidence of task diversity, the study examined whether the generated tasks introduced genuinely different functional requirements, providing a pedagogically meaningful basis for evaluating LLM-based exercise generation in introductory programming.

II. RELATED WORK

Large language models (LLMs) have become a central technology in both text generation and code-related educational support. Patil and Gudivada [1] reviewed the progression from task-specific natural language processing (NLP) architectures to transformer-based LLMs. They identified pretraining, transfer learning, in-context learning, model scale, and data coverage as key factors shaping model behavior. In computer science education, Raihan et al. [2] showed that the research literature is concentrated mainly in undergraduate and introductory programming contexts, where LLMs are studied for code generation, explanations, feedback, task generation, and assessment support. This educational context is directly relevant to generating beginner programming tasks, in which outputs must be technically correct, understandable to novices, and pedagogically aligned with the target concepts. Automated programming exercise generation is usually motivated by the high cost of manually preparing large numbers of suitable tasks. Frankford et al. [3] synthesized early work on LLM-based programming exercise generation and reported that multiple models can produce useful exercises, while also emphasizing the need for systematic evaluation and highlighting challenges such as quality assurance and the ease with which LLMs may solve LLM-generated tasks. Meißner et al. [4] compared ChatGPT with GPT-4, Bing AI Chat with GPT-4, Google Bard with LaMDA, and earlier GPT-3.5-turbo results for Java exercise generation. Their findings indicated that LLMs can initiate exercise designs across different topics, but manual refinement remains important, especially for more complex programming concepts and error-classification tasks. Moving from comparative evaluation to tool-supported generation, Ta et al. [5] proposed ExGen, a Visual Studio Code extension for on-demand generation of introductory Python exercises using GPT-3.5 Turbo and GPT-4 Turbo. Their framework defined a ready-to-use exercise as one with a clear problem statement, working reference solution, test cases, and an appropriate requested difficulty level. It used chained filtering checks such as compilation, unit testing, and difficulty verification. Zeng et al. [6] focused on open-source LLMs for generating programming exercises and showed that instruction tuning can improve selected models. The use of Evol-Instruct, an automatic instruction-evolution method for generating more complex instructions from simpler ones, and ACES (Autotelic Code Exploration via Semantic descriptors) is particularly relevant to diversity. Both methods support the construction of varied instruction data: Evol-Instruct increases complexity through iterative instruction rewriting. At the same time, ACES guides generation using semantic descriptors such as difficulty levels and programming concepts. Finally, Zou et al. [7] examined ChatGPT for algorithmic exercise generation and grading in C-oriented algorithmic education, finding strong sensibleness, readiness, and topicality, but identifying novelty as a weaker dimension. These approaches have demonstrated the potential to produce exercises with different surface characteristics, difficulty levels, and topical coverage. However, they have typically evaluated diversity indirectly, for example, through sensibleness, novelty, readiness, topicality, or accuracy, rather than by systematically analyzing the underlying computational operations required to solve the generated tasks. In parallel, many studies have examined diversity, creativity, and sampling in LLM outputs more directly. Bellemare-Pepin et al. [8] benchmarked LLMs against a large human cohort on the divergent association and creative writing tasks, using semantic-divergence measures. They found that prompt strategies and higher temperature can increase semantic divergence, but that the most creative human participants still outperform the strongest LLMs. Zhou et al. [9] formalized the diversity-quality trade-off in open-ended text generation by evaluating sampling methods, showing that parameter selection and adaptive truncation affect the balance between diversity and risk. Ságodi et al. [10] provided complementary evidence for code-related generation by releasing a program-synthesis dataset generated from nine open-source models across the Llama, Qwen, and DeepSeek families, using seven programming tasks and 100 temperature settings, repeated three times. In summary, prior work established that LLMs can support the generation of programming exercises, but it also exposed several gaps. First, many studies evaluated textual quality, functional correctness, novelty, creativity, or readiness separately rather than asking whether surface-level variation produced genuinely different task types. Second, diversity is operationalized differently across domains: embedding-based semantic distances in creative-language studies, recall-risk metrics in decoding research, and manual or search-based novelty assessment in programming-exercise studies. Third, temperature has been studied in open-ended creativity and program synthesis, but less is known about its effect on constrained beginner-level programming

exercises. The present study addressed these gaps by distinguishing textual diversity from functional diversity in LLM-generated beginner C array tasks and examining whether higher temperatures produced genuinely new functional task types across multiple cloud-based models.

III. MATERIALS AND METHODS

1. Experimental Setup

The study was designed to examine the extent of variation produced by cloud-based LLMs in beginner-level C programming tasks when the prompts were systematically controlled. Variation was evaluated at two levels: textual diversity, referring to differences in task wording, and functional diversity, referring to differences in the underlying programming operations required to solve the tasks. The experimental workflow is shown in Figure 1. For each model, prompts were submitted in four task categories and at four temperature settings. Each model–prompt–temperature condition was repeated 10 times, resulting in a total of 960 generated outputs: six models × four prompt categories × four temperature settings × ten repetitions—the generated tasks focused on introductory C programming with one-dimensional integer arrays. The prompts constrained the tasks to concepts typically covered in beginner programming courses, including loops, conditionals, and arrays. The requested output format was a JSON object containing only the task description, an example input, and an example output. This allowed the outputs to be parsed systematically and compared across models, prompt categories, and temperature settings.

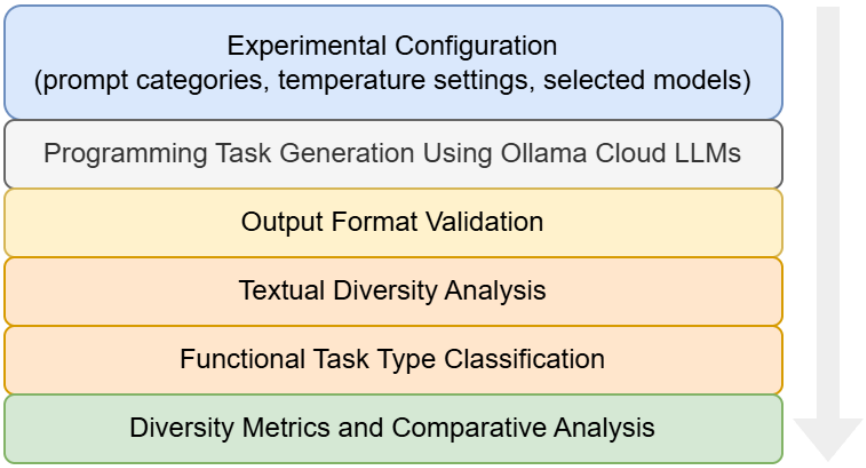


Figure 1. Experimental workflow for evaluating textual and functional diversity in LLM-generated beginner programming tasks.

2. Evaluated Models

Six Ollama cloud-based LLMs were evaluated in the experiment: DeepSeek V4 Pro, Qwen3 Coder Next, Kimi K2.6, GLM 5, Gemma 4, and Nemotron 3 nano. The selected models included both general-purpose and coding-oriented systems, as well as dense and mixture-of-experts architectures. Their Ollama model identifiers, providers, types, approximate sizes, and primary focus areas are summarized in Table 1.

Table 1. Characteristics of the evaluated LLMs

Ollama Model Identifier	Provider	Type	Approx. Size	Primary Focus
deepseek-v4-pro: cloud [11]	DeepSeek	MoE	1.6T (49B active)	Reasoning, coding
qwen3-coder-next: cloud [12]	Alibaba	MoE	80B (3B active)	Coding
kimi-k2.6:cloud [13], [14]	Moonshot AI	MoE	1.04T (32B active)	Coding, agents
glm-5:cloud [15]	Z.ai / Zhipu AI	MoE	744B (40B active)	Reasoning, coding
gemma4:31b-cloud [16]	Google	Dense	31B	General-purpose
nemotron-3-nano:30b-cloud [17]	NVIDIA	MoE	30B (3.5B active)	Reasoning, coding

The study used the Ollama cloud model variants because they provided a consistent access interface while representing different current LLM families. The comparison was not intended to benchmark general model quality but to examine how different models responded to the same task-generation prompts and temperature settings.

3. Prompt Categories and Temperature Settings

Four prompt categories were used to represent common types of beginner-level array-processing tasks: aggregation, search, modification, and two-array processing. The full prompts are shown in Table 2. Each prompt defined the expected input structure, output type, and allowed programming constructs. Additional restrictions were included to keep the generated tasks within the intended beginner-level scope.

Table 2. Prompts used for generating programming tasks

Task	Prompt
aggregation	Generate a beginner-level C programming task that processes all elements of an integer array to produce a single result. Input: n (1<=n<=100) and n integers. Output: one value derived from the array. Use only loops, arrays, and conditionals: no pointers, no dynamic memory, no sorting.
search	Generate a beginner-level C programming task that requires examining elements of an integer array to find specific information. Input: n (1<=n<=100), n integers, and optionally an additional value. Output: a result related to locating or identifying elements. Use only loops, arrays, and conditionals—no sorting or binary search.
modification	Generate a beginner-level C programming task that modifies elements of an integer array directly. Input: n (1<=n<=100) and n integers. Output: the modified array. Use only loops, arrays, and conditionals. No auxiliary arrays.
two_arrays	Generate a beginner-level C programming task that uses two integer arrays of equal length to produce a result. Input: n (1<=n<=100) and two arrays of size n. Output: a result based on both arrays. Use only loops, arrays, and conditionals—no sorting or advanced structures.

A common suffix was appended to every prompt to enforce a uniform output format. As shown in Table 3, the suffix required the task to be solvable in fewer than 20 lines of C code and instructed the model to return only valid JSON with exactly three keys: task, example_input, and example_output.

Table 3. Common prompt suffix used for output format constraints

The task must be solvable in under 20 lines of C code. Return ONLY valid JSON with EXACTLY these keys: task, example_input, example_output. Each value must be a single-line plain string. Example format: {"task":"...", "example_input":"...", "example_output":"..."}. Do not include any extra text.

Temperature was used as a sampling parameter. Four temperature values were tested: 0.1, 0.3, 0.6, and 0.9. These values were selected to cover low, moderate, and high sampling variability while keeping the generation conditions comparable across all models and prompt categories.

4. Diversity and Reliability Metrics

Two diversity measures were computed for each model–prompt–temperature condition. Textual diversity was defined as the number of distinct task descriptions generated. Two outputs were treated as textually identical only when the task description was identical. This metric captures surface-level variation in wording.

Functional diversity was defined as the number of distinct functional task types after semantically validated grouping. Outputs were assigned to the same functional group when they required the same input structure, the same main programming operation, and the same type of expected output, even if the wording, examples, or phrasing differed. For example, several tasks that asked students to count elements greater than the array's average were considered functionally equivalent, although their textual formulations differed. This distinction made it possible to examine whether greater textual variation corresponded to genuinely new programming tasks. Because fully manual grouping had proved time-consuming in an earlier, similar experiment [18], the present study used an LLM-assisted pre-labeling procedure to support the initial annotation of the 960 generated tasks. The processed CSV file containing the generated task descriptions was provided to ChatGPT together with functional-label examples from the earlier experiment. The resulting labels were used only as preliminary group assignments, which were then manually checked and corrected by one of the authors using the functional-equivalence criteria described above. Manual verification and correction were necessary because the LLM occasionally produced inaccurate, inconsistent, or overly generic labels. After this human validation step, the final functional-group labels were used to compute functional diversity. Textual and functional diversity counts were normalized by the number of generated outputs in the corresponding condition and are reported as percentages in the results.

In addition to diversity, output reliability was evaluated using three indicators of formatting and validity. The valid JSON rate measured the proportion of outputs that could be parsed as JSON with the required fields. When outputs were enclosed in markdown code block markers, these markers were removed before JSON parsing; the markdown block rate was reported separately to capture this formatting behavior. The markdown block rate, therefore, measured how often models returned JSON within a markdown code block, despite the instruction to avoid extra formatting. The truncation rate measured the proportion of incomplete outputs in which the response was cut off or did not contain a complete, usable JSON object.

For statistical analysis, two sets of two-way ANOVAs were conducted separately for textual and functional diversity. The first set examined the effects of model and temperature, while the second examined the effects of prompt type and temperature. Partial η^2 was reported as the effect-size measure.

IV. RESULTS

This section presents the results of the diversity and reliability analyses. The results are organized into three parts. First, overall textual and functional diversity are compared across the evaluated models. Second, the effects of model, temperature, and prompt category are examined using descriptive results and two-way ANOVAs. Third, output reliability is evaluated based on JSON validity, markdown code block usage, and truncation. The dataset underlying the analyses, including the generated model outputs, parsing and reliability indicators, and the manually validated functional grouping labels, is available on Zenodo [19].

1. Overall Textual and Functional Diversity Across Models

To make the derived metric explicit, let N denote the number of generated outputs, T the number of distinct task descriptions, and F the number of distinct manually validated functional groups. Textual diversity was calculated as T/N, functional diversity as F/N, and the functional-to-textual diversity ratio as F/T. The ratio, therefore, expresses how many distinct functional task types were obtained relative to the number of textually distinct task descriptions. A lower ratio indicates that many textually unique outputs still map onto the same underlying computational operation. The absolute counts used to calculate these measures are reported in Table 4. The models produced between 16 and 51 functional groups, compared with 84 to 141 distinct task descriptions.

Table 4. Absolute counts are used to calculate the textual and functional diversity across models

Ollama Model Name	**Generated outputs, N**	**Distinct task descriptions, T**	**Functional groups, F**
deepseek-v4-pro: cloud [11]	160	120	39
qwen3-coder-next: cloud [12]	160	93	27
kimi-k2.6:cloud [13]	160	113	48
glm-5:cloud [15]	160	118	40
gemma4:31b-cloud [16]	160	84	16
nemotron-3-nano:30b-cloud [17]	160	141	51

Figure 2 summarizes the overall textual diversity, functional diversity, and functional-to-textual diversity ratio across the six evaluated models. In all models, textual diversity was substantially higher than functional diversity. This indicates that the models frequently generated differently worded task descriptions that nevertheless corresponded to a smaller set of underlying computational operations.

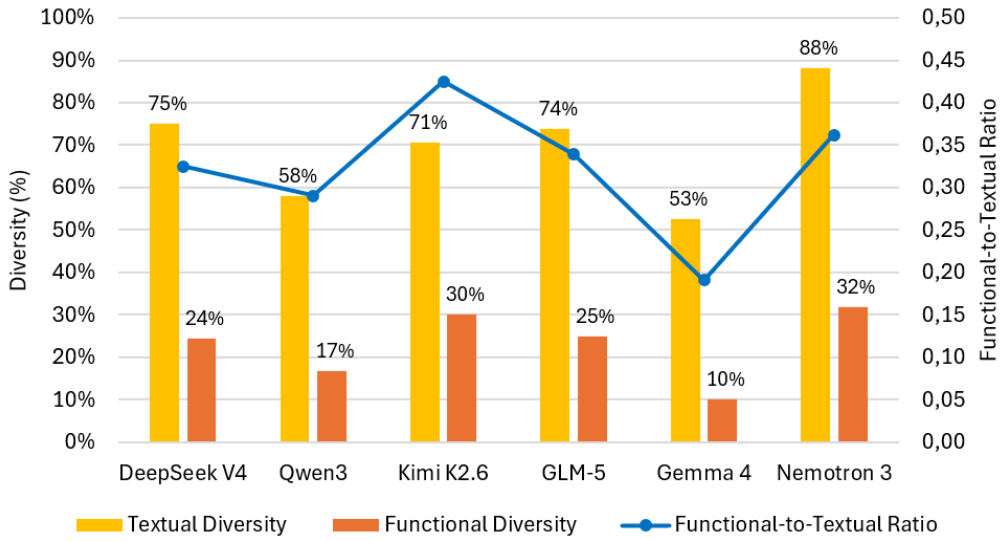


Figure 2. Overall textual diversity, functional diversity, and functional-to-textual diversity ratio across the evaluated LLMs

Nemotron 3 produced the highest overall textual diversity (88%) and also the highest functional diversity (32%). Kimi K2.6 showed the second-highest functional diversity (30%) and the highest functional-to-textual diversity ratio, suggesting that a relatively larger proportion of its textual variation corresponded to distinct functional task types. DeepSeek V4 and GLM-5 also produced high textual diversity, 75% and 74%, respectively, but their functional diversity remained considerably lower, at 24% and 25%. Qwen3 showed moderate textual diversity (58%) and functional diversity (17%). Gemma 4 had the lowest values on both measures, with 53% textual diversity and 10% functional diversity.

The functional-to-textual ratio further illustrates the difference between surface-level variation and functional variation. The ratio remained below 0.5 across all models, indicating that each model produced fewer distinct functional groups than distinct task descriptions. This pattern underscores the need to evaluate LLM-generated programming tasks not only for textual uniqueness but also for the underlying programming operations students are expected to perform.

Table 5 provides examples of textually different but functionally equivalent outputs. Although the three task descriptions use different wording, they all require the same operation: counting the array elements greater than the average of all elements.

Table 5. Examples of textually different but functionally equivalent programming tasks generated by DeepSeek V4 Pro at a temperature of 0.9

Write a C program that reads an integer n (1<=n<=100), then n integers, and outputs the count of elements that are strictly greater than the average of all elements (use integer division for the average).
Write a C program that reads an integer n (1<=n<=100), then reads n integers into an array. Find and print the count of elements that are strictly greater than the average of all elements (rounded down to the nearest integer).
Read n and an array of n integers. Count how many elements are greater than the average of all array elements. Output that counts.

In contrast to Table 5, Table 6 shows examples from the modification prompt category that were assigned to different functional groups because they require different modification operations.

Table 6. Examples of tasks from different functional groups in the modification category generated by DeepSeek V4 Pro at a temperature of 0.9

Read an integer n (1<=n<=100), then read n integers into an array. Modify the array in place: if an element is even, double it; if odd, add 1. Print the modified array separated by spaces.
Read an integer n, then an array of n integers. For each element, if it is positive and even, double it; if it is positive and odd, add 1; if it is negative, make it positive; if it is zero, leave it unchanged. Print the modified array separated by spaces.
Read an integer n (1<=n<=100), then read n integers into an array. Traverse the array and double each element that is positive and even, and set each negative element to zero. Leave other elements unchanged. Print the modified array on a single line, separated by spaces.

2. Effects of Model, Temperature, and Prompt Category on Textual and Functional Diversity

Figure 3 shows the effect of temperature on textual diversity across models. In general, textual diversity increased with temperature. This pattern was especially clear for DeepSeek V4, Qwen3, Kimi K2.6, and Nemotron 3, where the highest temperatures produced near-complete or complete textual uniqueness. DeepSeek V4 increased from 48% at a temperature of 0.1 to 100% at temperatures of 0.6 and 0.9. Qwen3 increased from 58% to 100%, while Kimi K2.6 increased from 53% to 100%. Nemotron 3 also reached 100% at a temperature of 0.9. GLM-5 started at a comparatively high level, even at 0.1, and reached 95% at 0.6 and 0.9. Gemma 4 showed the weakest increase, rising from 53% to 85%.

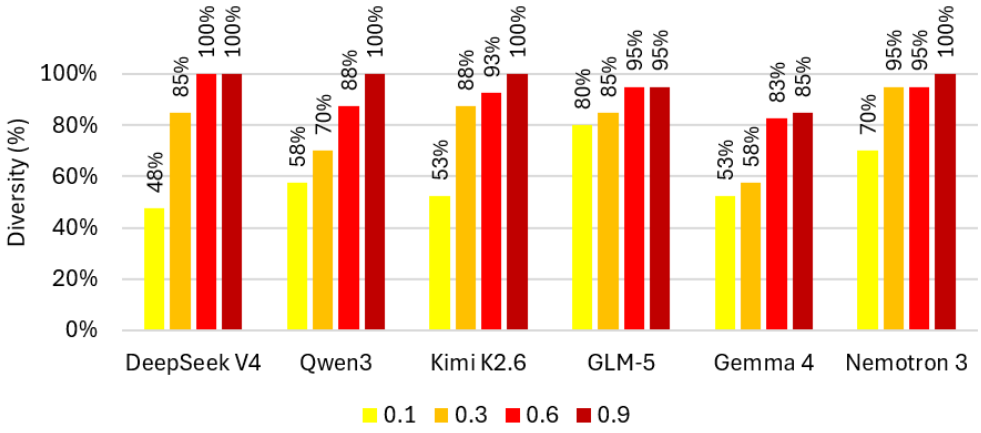


Figure 3. Effect of temperature on textual diversity across models

Figure 4 shows that functional diversity also tended to increase with temperature, but the effect was weaker and less uniform than for textual diversity. DeepSeek V4 increased from 28% to 68%, and Nemotron 3 showed the strongest increase, from 28% to 78%. Kimi K2.6 increased from 33% to 58%, while GLM-5 increased from 33% to 50%. Qwen3 showed only a modest increase, from 28% to 43%. Gemma 4 remained low across all temperature settings and showed no consistent upward trend.

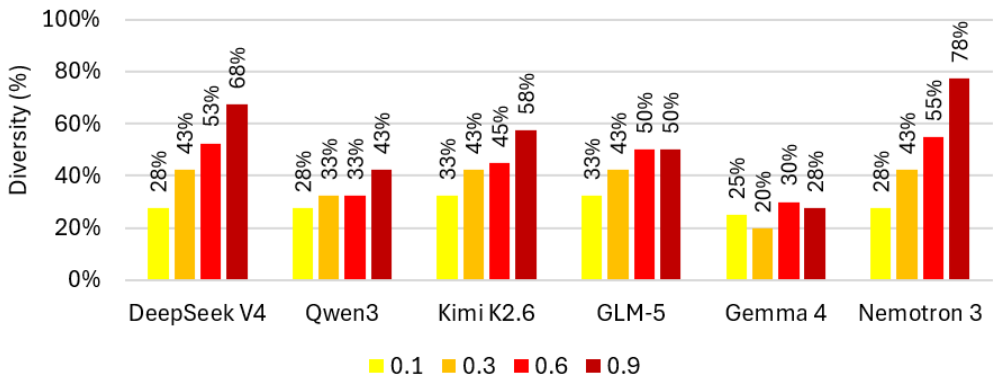


Figure 4. Effect of temperature on functional diversity across models

The two-way ANOVA for Model × Temperature is shown in Table 7. For textual diversity, both model and temperature had statistically significant effects. The effect of temperature was particularly strong, $F = 31.749$, $p < 0.001$, partial $\eta^2 = 0.570$, indicating that sampling temperature was the main driver of textual variation. The model effect was also significant, $F = 4.536$, $p = 0.001$, partial $\eta^2 = 0.240$. The Model × Temperature interaction was not statistically significant, $F = 1.233$, $p = 0.269$, suggesting that although models differed in overall diversity, the general effect of increasing temperature was broadly similar across models.

Table 7. Two-way ANOVA results (Model × Temperature)

Dependent Variable	**Factor**	**F**	**p**	**Partial η^2**
Textual Diversity	Model	4.536	0.001	0.240
Textual Diversity	Temperature	31.749	< 0.001	0.570
Textual Diversity	Model × Temperature	1.233	0.269	0.204
Functional Diversity	Model	3.427	0.008	0.192
Functional Diversity	Temperature	6.429	0.001	0.211
Functional Diversity	Model × Temperature	0.458	0.954	0.087

For functional diversity, both model and temperature were again statistically significant. The model effect was $F = 3.427$, $p = 0.008$, partial $\eta^2 = 0.192$, and the temperature effect was $F = 6.429$, $p = 0.001$, partial $\eta^2 = 0.211$. Compared with textual diversity, the temperature effect was smaller for functional diversity. This indicates that higher temperatures increased the likelihood of new wording more strongly than the likelihood of genuinely new functional task types. The Model × Temperature interaction was not significant, $F = 0.458$, $p = 0.954$.

Figure 5 compares diversity across prompt types. The two-array prompt category produced the highest textual diversity (78%), followed by search (76%), modification (66%), and aggregation (59%). However, the pattern for functional diversity was different. Modification tasks produced the highest functional diversity (40%), despite not having the highest textual diversity. Aggregation produced 15% functional diversity, two-array tasks 12%, and search tasks only 10%. Consequently, the modification category had the highest functional-to-textual ratio, indicating that textual variation in this category was more likely to correspond to genuinely different computational operations.

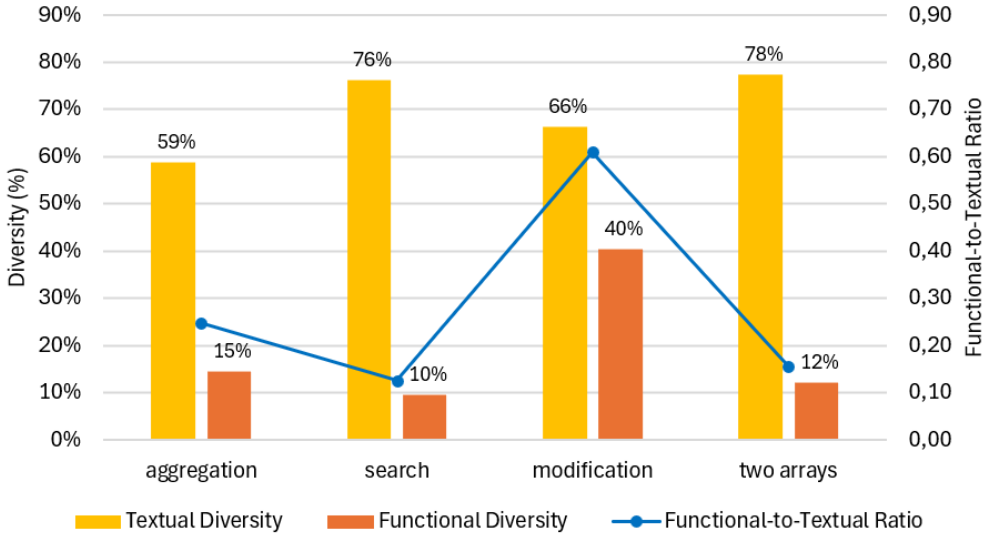


Figure 5. Overall textual diversity, functional diversity, and functional-to-textual diversity ratio across the prompt categories

The two-way ANOVA for Prompt × Temperature is shown in Table 8. For textual diversity, both prompt category and temperature were statistically significant. The prompt effect was $F = 5.601$, $p = 0.002$, partial $\eta^2 = 0.174$, while the temperature effect was stronger, $F = 28.854$, $p < 0.001$, partial $\eta^2 = 0.520$. The Prompt × Temperature interaction was not significant, $F = 0.672$, $p = 0.732$.

Table 8. Two-way ANOVA results (Prompt × Temperature)

Dependent Variable	**Factor**	**F**	**p**	**Partial η²**
Textual Diversity	Prompt	5.601	0.002	0.174
Textual Diversity	Temperature	28.854	< 0.001	0.520
Textual Diversity	Prompt × Temperature	0.672	0.732	0.070
Functional Diversity	Prompt	13.964	< 0.001	0.344
Functional Diversity	Temperature	8.324	< 0.001	0.238
Functional Diversity	Prompt × Temperature	0.269	0.981	0.029

For functional diversity, both prompt category and temperature were also significant. The prompt category effect was stronger than the temperature effect, $F = 13.964$, $p < 0.001$, partial $\eta^2 = 0.344$; $F = 8.324$, $p < 0.001$, partial $\eta^2 = 0.238$. This result indicated that the type of prompt had a greater effect on functional diversity than temperature did. In other words, increasing temperature generated more varied wording, but the prompt category more strongly determined whether the generated tasks represented different underlying programming operations. The Prompt × Temperature interaction was not statistically significant, $F = 0.269$, $p = 0.981$.

3. Reliability and Formatting Behavior

Table 9 summarizes the output reliability and formatting behavior of the evaluated models. Overall, the generated outputs were highly reliable in terms of JSON validity. Five models produced valid JSON in all cases, and Nemotron 3 did so in 99.4% of outputs. Truncation was rare, occurring only in Nemotron 3 at a rate of 0.6%.

The main reliability issue was not JSON validity, but unwanted markdown formatting. DeepSeek V4, Qwen3, and Nemotron 3 did not use markdown code blocks. Kimi K2.6 returned JSON inside markdown code blocks in 19.4% of the outputs. The issue was substantially more frequent for GLM-5 and Gemma 4, with markdown block rates of 68.1% and 84.4%, respectively. Although these outputs were still valid JSON after removing the surrounding markdown formatting, such behavior is relevant for automated educational pipelines because additional preprocessing may be required.

Table 9. Output reliability and formatting issues across evaluated LLMs

Ollama Model Name	**Markdown Block Rate (%)**	**Valid JSON Rate (%)**	**Truncation Rate (%)**
deepseek-v4-pro: cloud [11]	0%	100%	0%
qwen3-coder-next: cloud [12]	0%	100%	0%
kimi-k2.6:cloud [13]	19.4%	100%	0%
glm-5:cloud [15]	68.1%	100%	0%
gemma4:31b-cloud [16]	84.4%	100%	0%
nemotron-3-nano:30b-cloud [17]	0%	99.4%	0.6%

V. DISCUSSION

The findings highlight the importance of distinguishing between textual variation and functional diversity when evaluating LLM-generated programming tasks. Although higher sampling temperature generally increased the number of distinct task descriptions, this increase did not translate proportionally into new functional task types. In an educational context, this distinction is important because differently worded exercises may still require students to apply the same underlying computational operation. Therefore, textual uniqueness alone is not a sufficient indicator of pedagogically meaningful task diversity. If generated tasks are selected solely on the basis of textual uniqueness, a task set may appear varied yet offer limited conceptual coverage. Functional grouping or semantic filtering can therefore help ensure that generated exercises cover a broader range of programming concepts rather than only different phrasings of the same task. The results also suggest that temperature and prompt design affect different aspects of generation. Temperature mainly influenced surface-level variation, whereas the prompt category played a stronger role in determining the range of functional operations produced. This implies that simply increasing temperature is unlikely to be enough when the goal is to generate a broader set of practice tasks. More effective diversity control may require prompts that explicitly request different types of operations, constraints, or solution strategies. The reliability results indicate that structured-output prompting can be effective for producing machine-processable tasks, but formatting behavior remains model-dependent. Even when the generated content is valid, additional preprocessing may be needed if models wrap the output in unwanted markdown formatting. In practice, an educational platform could apply a simple filtering workflow: generate multiple candidate tasks, verify JSON validity, normalize formatting, cluster tasks by functional group, remove functional duplicates, and finally conduct expert review before selecting tasks for students.

VI. LIMITATIONS

This study has several limitations. First, the experiment focused only on beginner-level C programming tasks involving one-dimensional arrays. Therefore, the findings may not directly generalize to other programming languages, more advanced programming concepts, or different task formats. Second, the analysis included six Ollama cloud-based models and four prompt categories; other models, prompt templates, explicit functional-diversity instructions, or sampling parameters may produce different diversity patterns. Third, functional

diversity was based on semantically validated grouping. Although LLM-assisted pre-labeling was followed by manual checking and correction, the final grouping was verified by one author and still involved human judgment. Fourth, although the dataset contained 960 generated outputs, the effective sample size for statistical comparison was smaller because the diversity measures were calculated at the model, prompt category, and temperature levels rather than for each output. Each condition included ten repeated generations, which was sufficient to identify broad patterns, but smaller interaction effects should be interpreted with caution. Finally, the study evaluated the diversity of generated tasks and the reliability of outputs, but it did not test the pedagogical effectiveness of the generated tasks with students.

VII. CONCLUSION

This study evaluated textual diversity, functional diversity, and output reliability in LLM-generated beginner-level C programming tasks. The results demonstrated that textually distinct outputs often corresponded to the same underlying computational operation, indicating that textual uniqueness alone was insufficient to assess the pedagogical variety of generated exercises. The study contributes a distinction between surface-level task variation and functionally meaningful task variation, supported by semantically validated grouping of generated outputs. The findings suggested that effective LLM-based task generation should combine structured prompting with semantic validation rather than relying only on sampling parameters such as temperature. More specifically, prompt designs should explicitly target functional coverage rather than only asking for varied wording; for example, prompts can specify different operations, conditions, input structures, or solution strategies to support both textual variation and greater functional diversity. The evaluated models were able to produce structured programming tasks with high JSON validity. Still, the results also indicated the need for validation, filtering, and human supervision before such tasks are used in educational practice. Future work should extend the analysis to other programming topics, task formats, and empirical classroom evaluations.

ACKNOWLEDGEMENT

This work was supported by the national project KEGA 014TTU-4/2024, “Intelligent Animation-Simulation Models, Tools, and Environments for Deep Learning,” funded by the Ministry of Education of the Slovak Republic.

REFERENCES

[1] R. Patil and V. Gudivada, “A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs),” *Applied Sciences*, vol. 14, no. 5, p. 2074, Mar. 2024, doi: 10.3390/app14052074.

[2] N. Raihan, M. L. Siddiq, J. C. S. Santos, and M. Zampieri, “Large Language Models in Computer Science Education: A Systematic Literature Review,” in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, Pittsburgh, PA, USA: ACM, Feb. 2025, pp. 938–944. doi: 10.1145/3641554.3701863.

[3] E. Frankford, I. Höhn, C. Sauerwein, and R. Breu, “A Survey Study on the State of the Art of Programming Exercise Generation Using Large Language Models,” in *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*, Würzburg, Germany: IEEE, Jul. 2024, pp. 1–5. doi: 10.1109/CSEET62301.2024.10662990.

[4] N. Meißner, S. Speth, and S. Becker, “Automated Programming Exercise Generation in the Era of Large Language Models,” in *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*, Würzburg, Germany: IEEE, Jul. 2024, pp. 1–5. doi: 10.1109/CSEET62301.2024.10662984.



[5] N. B. D. Ta, H. G. P. Nguyen, and S. Gottipati, "Automatic generation of introductory programming exercises with large language models," *RPTEL*, vol. 21, p. 025, Sep. 2025, doi: 10.58459/rptel.2026.21025.

[6] G. Zeng, Q. Xue, and X. Lu, "Instruction Tuning with LLMs for Programming Exercise Generation," in *Web Information Systems and Applications*, vol. 14883, C. Jin, S. Yang, X. Shang, H. Wang, and Y. Zhang, Eds., in Lecture Notes in Computer Science, vol. 14883. , Singapore: Springer Nature Singapore, 2024, pp. 377–385. doi: 10.1007/978-981-97-7707-5_31.

[7] W. Zou, T.-T. Goh, H. Zhu, M. Liu, and B. Yang, "Algorithmic Learning: Assessing the Potential of Large Language Models (LLMs) for Automated Exercise Generation and Grading in Educational Settings," *International Journal of Human–Computer Interaction*, vol. 42, no. 2, pp. 1206–1223, Jan. 2026, doi: 10.1080/10447318.2025.2520931.

[8] A. Bellemare-Pepin *et al.*, "Divergent creativity in humans and large language models," *Sci Rep*, vol. 16, no. 1, p. 1279, Jan. 2026, doi: 10.1038/s41598-025-25157-3.

[9] Y. Zhou, M. Keuper, and M. Fritz, "Balancing Diversity and Risk in LLM Sampling: How to Select Your Method and Parameter for Open-Ended Text Generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria: Association for Computational Linguistics, 2025, pp. 26352–26365. doi: 10.18653/v1/2025.acl-long.1278.

[10] Z. Ságodi, I. Kolláth, P. Hegedűs, and R. Ferenc, "A Program Synthesis Dataset for LLM Temperature Analysis," *IEEE Access*, vol. 13, pp. 184022–184029, 2025, doi: 10.1109/ACCESS.2025.3625443.

[11] "deepseek-v4-pro: cloud." Accessed: May 29, 2026. [Online]. Available: https://ollama.com/deepseek-v4-pro:cloud

[12] "qwen3-coder-next: cloud." Accessed: May 29, 2026. [Online]. Available: https://ollama.com/qwen3-coder-next:cloud

[13] "kimi-k2.6:cloud." Accessed: May 29, 2026. [Online]. Available: https://ollama.com/kimi-k2.6:cloud

[14] "moonshotai/Kimi-K2.6 · Hugging Face." Accessed: Jun. 29, 2026. [Online]. Available: https://huggingface.co/moonshotai/Kimi-K2.6

[15] "glm-5:cloud." Accessed: May 29, 2026. [Online]. Available: https://ollama.com/glm-5:cloud

[16] "gemma4:31b-cloud." Accessed: May 29, 2026. [Online]. Available: https://ollama.com/gemma4:31b-cloud

[17] "nemotron-3-nano:30b-cloud." Accessed: May 29, 2026. [Online]. Available: https://ollama.com/nemotron-3-nano:30b-cloud

[18] L. Végh and D. Demeter, "Evaluating the Diversity of LLM-Generated Beginner C Programming Tasks Using Local Ollama Models," *IJANSER*, vol. 10, no. 4, pp. 64–72, Apr. 2026.

[19] L. Végh and D. Demeter, "Dataset for Textual and Functional Diversity Analysis of LLM-Generated Programming Exercises." Zenodo, Jul. 01, 2026. doi: 10.5281/zenodo.21093505.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Impact of AI Coding Agents on Software Engineering with Junior Developers: An Experiment in Situ

Sebastian Steger
Department of Computer Science
Technische Hochschule Mannheim
Mannheim, Germany
s.steger@th-mannheim.de
ORCID: 0009-0007-7480-3219

Kerstin Fink
Department of Social Work
Technische Hochschule Mannheim
Mannheim, Germany
k.fink@th-mannheim.de
ORCID: 0009-0007-6294-1518

Lukas Nock
Department of Social Work
Technische Hochschule Mannheim
Mannheim, Germany
l.nock@th-mannheim.de
ORCID: 0009-0004-3513-5666

Sven-Gunnar Klaus
Department of Computer Science
Technische Hochschule Mannheim
Mannheim, Germany
s.klaus@th-mannheim.de
ORCID: 0000-0003-4750-003X

Abstract
AI coding agents like Claude Code, Codex, and GitHub Copilot are currently disrupting software engineering practices. As today's junior developers will constitute tomorrow's workforce, understanding the impact of these tools on them and on the software they create is important. In this paper, we compare code quality, productivity, code comprehension, and team dynamics among three teams of four computer science students working on the same 10-week software engineering project, with only one team given access to an AI coding agent. Static code analysis, a post-study survey, and interviews, followed by a thematic analysis, were conducted. The results show similar code quality and overall productivity, but reduced codebase comprehension and a negative impact on the team's technical and social learning processes when using the AI coding agent. These findings highlight the tension between the potential immediate benefits of using AI coding agents and a potential long-term decline in the development of the technical and social skills needed by the future software engineering workforce.

Keywords: AI coding agents, junior developers, code quality, development productivity, team dynamics

I. INTRODUCTION

The rise of AI coding agents over the last few years has been transforming *Classical Software Engineering* (SE 1.0) into AI-Assisted Software Engineering (SE 2.0), and ultimately into *Agentic Software Engineering* (SE 3.0) [1]. Whereas in SE 1.0, Software was hand-coded without AI assistance, in SE 2.0, tools like Cursor, Windsurf, Gemini CLI, and GitHub Copilot are integrated into the IDE and have access to the codebase to assist the developer with the creation of entire classes, functions, tests, and boilerplate code in response to natural language prompts. In SE 3.0, AI coding agents like Devin, Claude Code, Codex, Jules, and GitHub Copilot Coding Agent autonomously plan changes, execute tools, make changes, run tests, and submit pull requests on a task level, possibly working in a remote environment. The interaction between the human and the agent is predominantly conducted in natural language via a chat interface or as part of a developer platform like GitHub.
Ongoing research is trying to determine the impact of that transformation on software quality (e.g., [2], [3]), development productivity (e.g., [4], [5], [6]), and team dynamics (e.g., [7]). Moreover, the role of the human

developer shifts from writing code manually (SE 1.0) to reviewing code and writing code collaboratively with an AI assistant (SE 2.0), and finally to defining goals and constraints as well as reviewing final changes (SE 3.0) [1]. These tasks, which were previously reserved only for the most experienced developers, now need to be handled by developers of all career levels - including junior developers – because simpler tasks such as writing boilerplate code and conducting basic tests can now be handled by AI coding agents [8]. However, the impact of junior developers' use of AI coding agents on software engineering remains largely unexplored.
This paper provides initial insights by comparing code quality, productivity, code comprehension, and team dynamics across three computer science student teams working on the same 10-week software engineering project. One team was equipped with an AI coding agent (SE 3.0), while the other two teams were restricted to using general-purpose chatbots such as ChatGPT (SE 1.0 + chatbot).

II. RELATED WORK

Impact on Code Quality. Azeem et al. compare 300 Python programs written by GPT-4 with functionally equivalent programs created by an experienced human developer and find that AI-generated code was 28% longer, with 21% higher cyclomatic complexity, while being easier to read [2]. Asfandyar et al. compare 107 programs written by human experts with equivalent Claude 3 and GPT-4 generated programs to find 24% greater cyclomatic complexity, 2.1 times greater code duplication, 3 times more OWASP top 10 vulnerabilities, and a 15-40% slower execution time with 25% more memory consumption in AI-generated code [3].
Impact on Productivity. In 2025, Cui et al. randomly assigned an AI coding assistant to 4867 developers in their daily work to find that the number of completed tasks increased by 26.08%, with larger gains for junior developers [4]. In a 2025 randomized controlled trial, Becker et al. had 16 experienced developers complete 246 tasks and found a 19% slowdown in completion time when using AI tools [5]. In a follow-up study [6] with 827 tasks and 57 developers, the authors reported the opposite result, with a 6.4% speedup in completion time when using AI tools.
Impact on Code Comprehension. In 2026, an experimental study by Shen & Tamkin [9] showed that while AI assistance can increase productivity, especially through task delegation, excessive reliance on AI impairs skill formation, conceptual understanding, and debugging abilities.
Impact on Team Dynamics. A 2023 literature review by Zercher et al. [7] indicated that group processes such as communication and coordination are less effective in team-AI collaboration than in human groups.
Due to the rapid advancement of AI in software engineering, the conclusions drawn in those studies can only be applied with caution to current agentic software engineering processes (SE 3.0). Furthermore, the presented studies focus on experienced software developers [2],[3], small tasks [2],[3],[4],[5], or coding with the help of a simple chatbot [2],[3] or AI-assisted coding (SE 2.0)[4],[5]. To the best of our knowledge, the present study is the first in situ experiment to comprehensively assess the impact of AI coding agents (SE 3.0) on software engineering among junior developers.

III. METHODS

1. Project

The experiment was conducted as part of the *Software Development Project (SEP)*, a practice-oriented university course in which small student teams are tasked with developing a near-product-quality software solution *for real-world clients in* industry, largely on their own. The project under study took place over 10 weeks, from October 10th to December 17th in 2025. The company NFON served as an industry client. The goal was to create a *Multi-Repo AI Documentation Generator Portal* with a web interface similar to Google Code Wiki.



2. Team Composition and Experimental Setup

To assemble the project teams, we implemented a structured, data-driven approach using a self-assessment questionnaire developed in collaboration with Reutlingen University. Students answered eight questions covering various soft skills and traits. Data analysis methods were then used to cluster the 12 participating students into three balanced teams of four students each. For validation purposes, their grades in previously taken programming and software engineering-related classes were analyzed, revealing a similar distribution across all teams. Each team was provided with an empty Git repository on GitHub. One of the three teams (referred to as the experimental team) was randomly selected and granted access to GitHub Copilot for auto-code completion and chat-first development. Furthermore, the GitHub coding agent was activated on their repository to enable remote, autonomous creation of pull requests. Apart from a brief demonstration of the coding agent capabilities, no training was provided. The other two teams served as control groups, A and B, and were instructed not to use specialized AI coding tools but were allowed to use chatbots like ChatGPT for consultation and code snippet creation.

3. Codebase Analysis

After the implementation period, metrics were extracted from the source code used as indicators for the code quality attributes maintainability, security, and reliability. The size of the codebase was determined by measuring the effective lines of code with the Linux tool *cloc*. The results were grouped by application layer (backend and frontend) and by purpose (testing and production).
To measure code quality attributes, the TIOBE software quality framework, TiCS 2026.1 [10], was applied to the three codebases. It performs static code analysis and combines it with code coverage data from automated tests. Based on that, it computes a percentage score for the quality indicators: code security, code coverage, fan-out, code duplication, cyclomatic complexity, coding standards, and interpreter warnings. Those values are then combined into an overall code quality score.

4. Productivity Assessment

Productivity is defined as the total output divided by the total input. As all teams were following the same project goal, the total output is not affected by the type of goal, but only by the quality with which it is achieved. To measure that, we use the average satisfaction from both major stakeholders, the industry client, and the responsible professors. Both stakeholders assigned each team a score/grade based on the external and internal quality of the delivered product and the presentation of the results. Those numbers have been normalized to percentages, with 0% indicating total failure to meet requirements and 100% indicating perfect execution and stakeholder satisfaction. The total input is defined as the amount of time spent on the project. Therefore, we asked each participant once a week to report their individual number. In addition to that quantitative assessment, we also asked the participants about their own perceived progress relative to their expectations on a Likert scale from *a lot slower* to *a lot faster*.

5. Codebase Comprehension Assessment

The familiarity with the source code was assessed both qualitatively and quantitatively. For the former, in a post-project survey, all participants rated their familiarity with the part of the codebase they were responsible for and with the entire team's codebase on a Likert scale. For the latter, a one-hour interview was conducted with each team two days after the project ended. During the interview, the team members were asked to collaboratively create a static building-block architectural diagram (building blocks and their dependencies) at two

levels of abstraction (level 1: frontend/backend/database; level 2: high-level modules within each level-1 building block) without consulting the source code. Inspired by [11], we then compared them to the actual architecture extracted from the code using the graph edit distance (GED), which corresponds to the minimum number of modifications required to convert one diagram into another. The number of elements in the actual architecture diagram was then used to normalize the GED.

6. Team Dynamics Assessment

To investigate the effects of AI agent use on group dynamics and learning processes within student teams, a mixed-methods approach was employed, combining quantitative and qualitative social research methods. In a first step, all project groups completed a short, standardized questionnaire at the conclusion of the project phase. The survey assessed perceptions of interaction quality [12] and information processing within the teams [13]. Comparisons between the experimental and control groups were conducted using the scales *team processes* (Cronbach's $\alpha = .915$) and *information processing* (Cronbach's $\alpha = .784$), which were constructed from the relevant questionnaire items. In a second step, semi-structured interviews were conducted with members of both the experimental and control groups to capture participants' subjective perspectives on collaboration, process quality, and project outcomes. Attention was paid to differences in task distribution, decision-making, and conflict resolution, as well as to the role of the AI agent within these processes. The interview material was analyzed using thematic analysis [14], a qualitative method for identifying, examining, and interpreting recurring patterns in the data. The transcripts were coded inductively, and the resulting codes were subsequently grouped into overarching themes.

IV. RESULTS

1. AI Usage

In the post-study survey, participants were asked how often they had used each type of AI tool (none, ChatGPT-like, AI code completion, chat first, remote agent) for programming-related tasks. We mapped that to the software engineering paradigms described in Section I. The distribution of the experimental group's AI usage was 27% for SE 3.0 and 37% for both SE 2.0 and SE 1.0. Control Group A (B)'s values were 0% (9%), 25% (17%), and 75% (74%). Despite the requirement that those teams use only ChatGPT-like AI tools, they couldn't entirely resist using more advanced ones. Control Group B used the full-stack AI development platform Lovable for frontend development. Therefore, we excluded the frontend from the codebase analysis. We also asked the participants in the experimental group which tasks they had used the remote coding agent for and how satisfied they were with the results. The most common mentions were the implementation of entire features, followed by small changes, documentation, bug fixes, refactoring, and test creation. For DevOps-related tasks, code reviews, requirements engineering, and architectural work, usage of the coding agent played a minor role. According to the participants, 52% of

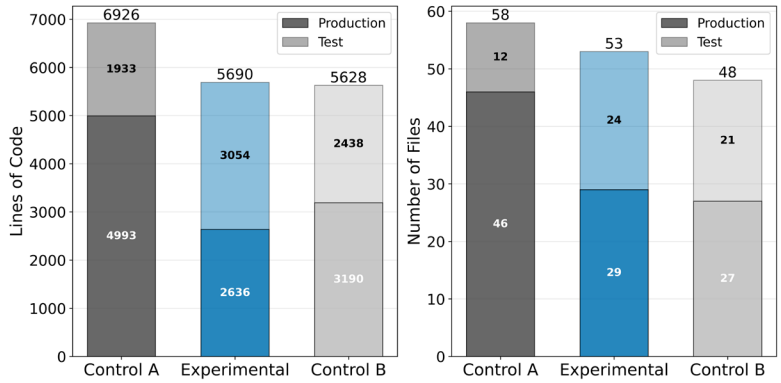


Figure 1. Codebase size of the production and test backend code in Lines of Code (left) and Number of Files (right)

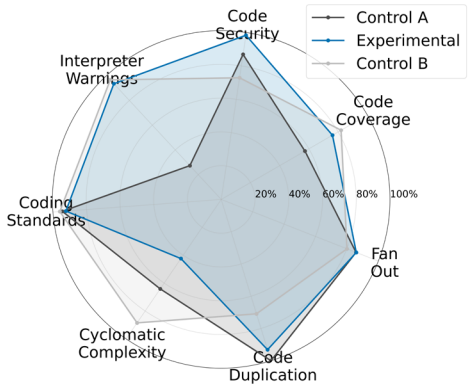


Figure 2. Contributing factors of the TIOBE Quality Indicator from worst (center) to best (outside)

The PRs have been merged without any changes after review. For 24% (9%), the agent (a human developer) improved the solution before merging, whereas in 10% of cases, the PR was discarded. The remaining 5% have been merged without any code review (vibe coding).

2. Codebase Analysis

With 2636 LoC, the production backend codebase of the experimental group was the smallest among the control groups A and B, at 4993 and 3190 LoC, respectively. In turn, the size of the test code was the largest, at 3054 LoC, versus 1933 (2438) LoC for control group A (B). Detailed numbers are shown in Fig. 1.
The TIOBE TiCS static code analysis resulted in an overall code quality score of 86% for the experimental group and 74% (88%) for control group A (B), respectively. The individual contributing factors are depicted in Fig. 2. The remainder of this section addresses each of them.

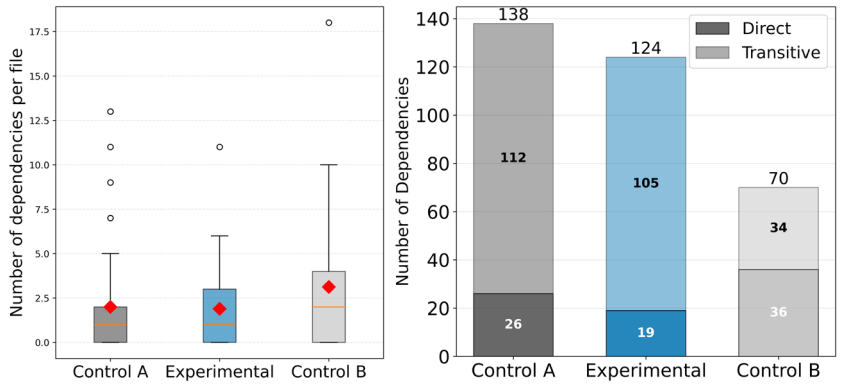


Figure 3. Fanout statistics. Left: statistics on internal dependencies (25, 50, 75th percentile, 1.5 IQR outliers, and mean (red diamond)). Right: Number of direct and transitive dependencies on 3rd party libraries.

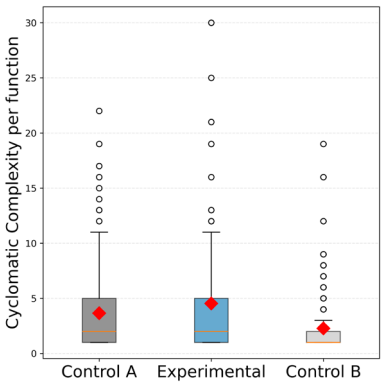


Figure 4. Function Cyclomatic Complexity statistics (25, 50,

75th percentile, 1.5 IQR outliers, and mean (red diamond)).

The **Code Security** analysis revealed some violations across the codebases, such as plaintext secrets, missing JWT signature validation, and *CORS policies set to allow all origins*. With an overall security score of 98%, the experimental group had the fewest serious violations, followed by control group A (87%) and control group B (73%).
With a statement **Code Coverage** of 67%, the experimental group achieved a value similar to that of the control group B (66%). Control group A had a minor statement coverage of 33%.

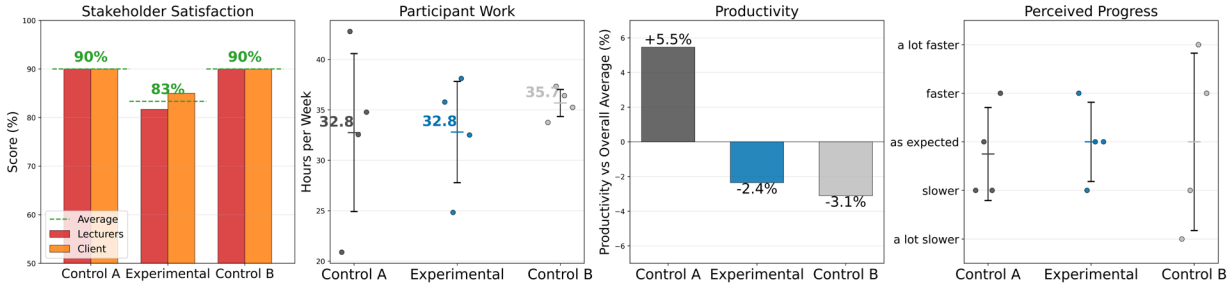


Figure 5. stakeholder satisfaction, participant work, relative productivity, and perceived progress by team

As a measure for coupling, internal Fan Out, i.e., the number of other source code files referenced by each file, was 1.9 on average for the experimental group and 2.0 (3.1) for control group A (B). The experimental group's requirements.txt has 19 dependencies on 3rd-party libraries (external fan-out), whereas control group A (B) has 26 (36) dependencies. The statistics are depicted in Fig. 3.

For control group B, 364 duplicated tokens were found, whereas no Code Duplication was found for the experimental group and control group A.

With an average Cyclomatic Complexity (CC) of 4.53, the experimental group's code is more complex than the control groups' codes, with CCs of 3.65 and 2.27, which may indicate reduced maintainability and increased error-proneness. The statistics are shown in Fig. 4.

The experimental group had 0 Level 1 Coding Standards violations, whereas the control group had 2. For Level 2, the values were 32 for the experimental group and 71 (22) for control group A (B).

Control group A Interpreter Warnings were worse than those of the other groups. They had 143 violations: Name may be undefined, or defined from star imports; 33 Module imported but unused; 4 from module import * used; and 3 Local variable name is assigned but never used [15].

3. Productivity

The measured and perceived productivity, along with their components, are depicted in Fig. 5. All teams completed the project. At 83%, the average stakeholder satisfaction with the experimental team's performance was lower than that of the control groups, which each scored 90%. There was almost no difference in the ratings of the university lecturers compared to those of the industry client. On average, the participants in the experimental team spent 32.8 ± 5.0 h working on the project over those 10 weeks, similar to control team A at 32.8 ± 7.8 h

per week, but less than control team B at 35.7 ± 1.3 h per week. The resulting productivity of the experimental group was 2.54 project goals·score/h, placing it between control group B (2.52) and A (2.74). Therefore, neither an overall performance boost nor a performance degradation was observed in the experimental group. Because all three teams rated their own progress *as expected* on average, no differences were evident in the self-assessments either.

4. Code Comprehension

In the post-study survey, the experimental group reported lower comprehension of the part of the codebase they were responsible for. Here, the ratings were between medium and good for the experimental group, versus medium and very good for the control group. The self-assessment difference was more pronounced for ratings between very poor and medium, rather than between poor and good. This self-assessment is backed up by the interviews conducted. The expert interviewer perceived greater uncertainty in the experimental group when creating class and sequence diagrams for the system being created. For the backend, the normalized graph edit distance between the memorized and actual building block diagram was highest (0.85) for the experimental group, compared with 0.77 and 0.50 for the control groups A and B, indicating lower comprehension of the codebase. Detailed numbers can be found in Fig. 6Figure 6.

5. Team Dynamics

To examine whether the experimental group differed significantly from the control group in their evaluation of team processes and information processing, a Mann-Whitney U test was conducted. The findings indicate that the experimental group rated both their team processes ($U = 3.000$; $Z = -2.084$; $p < .05$) and their information processing ($U = 2.000$; $Z = -2.294$; $p < .05$) significantly lower than the control groups. The underlying causes of these differences and the specific role of the AI agent initially remained unclear. In this regard, the subsequent thematic analysis of the interview data identified two contrasting themes: first, the AI agent as a facilitator of project success, and second, the AI agent as an impediment to learning processes.

Theme 1: The AI agent as facilitator for project success

Members of the experimental group described the AI agent as a substantial source of support and a considerable time-saving resource during the project. The agent was used extensively and enabled the rapid generation of large amounts of code, which could almost always be adopted after only brief quality checks. Participants further agreed that, given the existing difficulties within the team, the project would likely not have been completed successfully without the AI agent's support. Consequently, the agent was perceived not merely as an

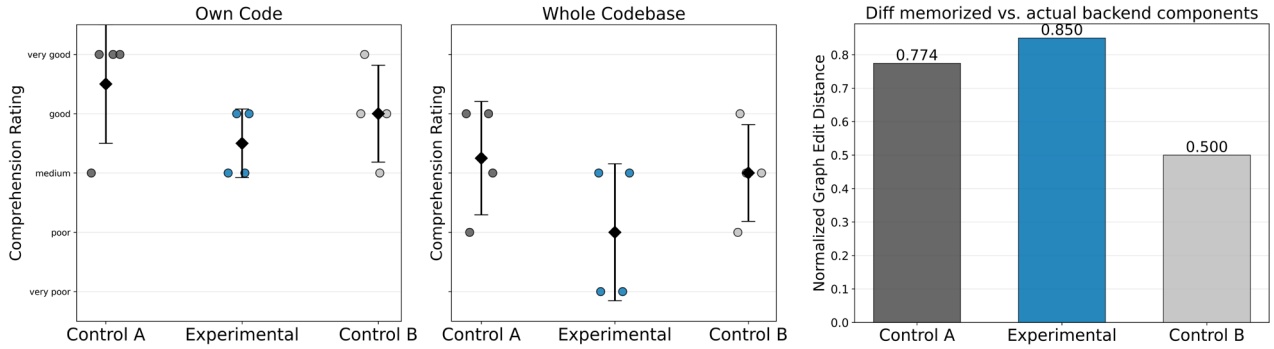


Figure 6. left: perceived code comprehension per team (scatter + mean ± SD); right: difference between memorized and actual backend.

efficiency-enhancing tool but as a decisive contributor to the project's successful completion. This perception was symbolically reflected in the team's assigning the agent a nickname and referring to it as their "*best colleague*".

Theme 2: The AI agent as an obstacle to learning processes

At the same time, the experimental group rated their code comprehension lower than that of the control group (See section 4.4). Within this team, tasks and problem-solving were increasingly delegated to the AI agent. Although participants initially attempted to solve problems on their own, this process was soon perceived as too time-consuming. Despite being aware of the potential negative implications for learning, students continued to rely heavily on the agent. Some participants even reported experiencing little or no improvement in their software engineering abilities after the project's completion. This form of *skill skipping* [16], in which learning opportunities remain untapped, is critically reflected upon by the students concerned: *"It [the AI agent] took care of everything for you, and you didn't have to grow through your problem*s."

Naturally, conflicts arose in all teams – but they were significantly more pronounced in the experimental group. Team members reported severe communication and coordination problems, as well as an overall negative team atmosphere. In some cases, it was unclear which individual members were actually working on or which tasks were next, leading to an uneven distribution of work, a blurring of responsibilities, and additional frustration. Initial attempts to improve processes and the team atmosphere increasingly gave way to a *turn a blind eye and carry on* mentality. Although the respondents themselves did not see the AI agent as the cause of the conflicts, its use reduced the need for close coordination and, in turn, the need to address existing problems. One member summed it up as follows: "*Okay, let's just swallow all the negativity and get on with it!*" The result was a working group characterized by weak social cohesion and limited personal identification with the project outcome.

V. DISCUSSION

Even though the experimental team had access to and was encouraged to use a state-of-the-art coding agent, this agent wasn't always the first choice. Instead, AI-assisted SE and even classical SE were used. We attribute this partly to the unfamiliarity with the process. When used, the junior developers were satisfied with the results in most cases, and trust grew to the point that they occasionally accepted them unquestioningly.

The codebase analysis revealed no clear difference between development with and without an AI coding agent. The differences in *coding standard violations* were minor, while *codebase size*, *code coverage*, dependencies, code duplication, and interpreter warnings all remained within the control group ranges. In contrast, the codebase created with assistance from the AI coding agent showed higher code security but potentially reduced maintainability, as indicated by its higher cyclomatic complexity. Despite prior studies reporting productivity gains from AI use [4], [6], our experiment does not confirm these findings. The team using the AI agent neither produced more software of better quality nor spent less time on the project than the other two teams. This is not necessarily a contradiction, because instead of looking at task-level productivity, we measured overall productivity over 10 weeks, including management tasks such as preparing presentations, team communication, and writing up documents, in addition to core software development-related tasks. This pattern resembles the "productivity paradox" observed in the 1980s, when the introduction of IT was not associated with higher productivity [17]. In our experiment, the exact reasons remain unclear.

Code Quality	Productivity	Code Comprehension	Learning Process
o	o	-	-

The findings on code comprehension and team dynamics highlight a tension between the subjectively perceived productivity gains from using AI agents and the simultaneous erosion of both technical and social learning processes. Within the experimental group, impairments in Learning became evident not only in participants' self-assessments of their programming competencies, but also in their objectively measured understanding of the generated code (see section 4.4).

Moreover, the results suggest that social learning processes within the teams were likewise inhibited. The AI agent made it easier for participants to avoid both intensive technical engagement with the code and the need to resolve team conflicts. At the same time, signs of alienation from the project output became apparent. Members of the experimental group identified less strongly with their final product than participants in the control groups. Together, the findings suggest that the use of AI agents in student project teams should be approached critically when the primary objective is competence development and learning rather than the efficient production of results alone. Or, as one member of the experimental group put it: *"Do I want to deliver a result in a project, or do I want to develop myself further?"*

Strengths of this work include the experimental and control groups having the same project goals and a balanced team composition, making the results inherently comparable. Furthermore, the study provides a holistic assessment of the impact on software engineering and was conducted in a natural setting (a student project) rather than in an isolated lab. The main limitation of this study is the small number of teams participating. The measured differences/similarities cannot be attributed solely to the use of the AI coding agent and may be superimposed by other influencing factors. Other limitations are the exclusion of the frontend code from the code quality analysis and the rapidly outdated AI coding agent used for studies like this.

VI. CONCLUSION AND FUTURE WORK

In this paper, we assessed the impact of AI coding agents on software engineering with junior developers. In a 10-week in-situ experiment, we observed the overall code quality, productivity, code comprehension, and team dynamics of three teams, each consisting of four junior developers working independently on the same project, with only one team given access to an AI coding agent. As summarized in Table 1, we observed no significant impact on code quality or overall productivity in that experiment. However, when analyzing code comprehension and team dynamics, we found that the AI agent facilitates project success by taking over tasks but, at the same time, removes responsibility, self-confidence, learning opportunities, and identification with the project outcome from team members. Together with the reported reduction in lower-level tasks [2], the hindered career advancement opportunities through the use of AI agents, as suggested by our experimental results, may reinforce the already ongoing decline in demand for junior software developers.

Future work includes repeating the experiment to increase the reliability of the results, as well as variations such as upstream training of participants in current AI software engineering practices and the incorporation of senior developers into development teams. Furthermore, best practices for the career advancement of junior software developers need to be adapted to the rise of agentic software engineering.

ACKNOWLEDGEMENT



The authors would like to thank Tim Hendriks and Jana Richter (NFON) for defining the project goals, Rob Goud and Paul Jansen (TIOBE) for supporting the code quality analysis, as well as Semjon Mössinger (WPS), Robert Glaser (Exxeta), and Jörg Müller (INNOQ) for the discussions during the design phase of the experiment.

REFERENCES

[1]. H. Li et al., "The Rise of AI Teammates in Software Engineering (SE)3.0: How Autonomous Coding Agents Are Reshaping Software Engineering." 2025. [Online]. doi: 10.48550/arXiv.2507.15003.

[2]. S. Azeem et al., "AI vs. Human Programmers: Complexity and Performance in Code Generation," VAWKUM Trans. Comput. Sci., vol. 13, pp. 201–216, May 2025, doi: 10.21015/vtcs.v13i1.2043.

[3]. Asfandyar, M. W. Akhtar, and H. S. Khalil, "Evaluating the Impact of AI Code Generated on Software Quality and Observed Study of Maintainability, Security and Performance," Spectrum Eng. Sci., vol. 3, no. 12, pp. 1379–1389, Dec. 2025.

[4]. Z. Cui et al., "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers," SSRN, Aug. 2025. doi: 10.2139/ssrn.4945566.

[5]. J. Becker et al., "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.", METR, Jul. 2025.

[6]. J. Becker et al., "We are Changing our Developer Productivity Experiment Design." Feb. 2026. [Online]. Available: https://metr.org/blog/2026-02-24-uplift-update/ (accessed on May 28th, 2026).

[7]. D. Zercher et al., "When AI joins the Team: A Literature Review on Intragroup Processes and their Effect on Team Performance in Team-AI-Collaboration", ECIS 2023. Research Papers. 307. Nov. 2023.

[8]. S. Panyam et al., "How AI Agents Are Transforming Software Engineering and the Future of Product Development," Computer, vol. 58, no. 05, pp. 71–77, May 2025, doi: 10.1109/MC.2024.3488378.

[9]. J.H. Shen and A. Tamkin, "How AI Impacts Skill Formation," arXiv preprint arXiv:2601.20245, 2026. [Online]. doi: 10.48550/arXiv.2601.20245.

[10]. P. Jansen, "The TIOBE Quality Indicator: A pragmatic way of measuring code quality," TIOBE, Technical Report TIOBE–20120718.1, Oct. 2025.

[11]. P. Čech, "Matching UML Class Models using Graph Edit Distance," Expert Systems with Applications, vol. 130, Apr. 2019, doi: 10.1016/j.eswa.2019.04.008.

[12]. R. Wageman, J.R. Hackman, and E. Lehman, "Team Diagnostic Survey: Development of an Instrument," The Journal of Applied Behavioral Science, vol. 41, no. 4, pp. 373–398, Dec. 2005.

[13]. J.A. Fischer and H. Hüttermann, "Fragebogen zur Messung von Informationsverarbeitung im Team (FIT)," Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS), 2021.

[14]. V. Braun and V. Clarke, "Using thematic analysis in psychology," Qualitative Research in Psychology, vol. 3, no. 2, pp. 77–101, 2006, doi: 10.1191/1478088706qp063oa.

[15]. G. McConnaughey, "Flake8 Rules." 2024. [Online]. Available: https://www.flake8rules.com/ (accessed on May 28th, 2026).

[16]. F. Nuxoll, "Skill Skipping. Was Lehrpersonen tun können, damit KI nicht Lernprozesse verhindert", Lehren & Lernen, vol. 50, no. 6, pp. 11–12, 2024.

[17]. E. Brynjolfsson, "The productivity paradox of information technology," Commun. ACM, vol. 36, no. 12, pp. 66–77, Dec. 1993, doi: 10.1145/163298.163309.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

HID Hardening vs. FIDO2: A Security-versus-Security Trade-off

Dr. Emin Huseynov
Institute of Data Science and Artificial Intelligence
Azerbaijan Technical University
Baku, Azerbaijan
emin@huseynov.com
ORCID: 0000-0003-4198-5400

Abstract

Problem. Human Interface Device (HID) keystroke-injection attacks ("BadUSB" / "Rubber Ducky") exploit the implicit trust operating systems grant to keyboards, and are now trivial to mount with hardware such as the US$4 Raspberry Pi Pico. On Windows, the standard defense is to block the installation of HID-class devices via Group Policy. Finding. We show this defense carries an overlooked cost: because FIDO2/WebAuthn security keys use the same USB HID transport (CTAPHID), a policy strict enough to block a Pico can also disable phishing-resistant authentication—a security-versus-security conflict, not merely security-versus-usability. Recommendation. Because CTAPHID occupies a dedicated FIDO usage page (0xF1D0), distinct from the keyboard usage page, injection-capable keyboards can be blocked while security keys remain reachable. We present a layered, descriptor-aware policy and validate it on Windows 11 (24H2). Across a four-case test, a Pico is blocked, while two vendors' FIDO2 keys still enroll and authenticate under the same policy. We further show that hardware-identifier allowlisting is necessary but insufficient—the Pico can spoof identifiers—thereby motivating attestation-based authorization.

Keywords: BadUSB; HID injection; Raspberry Pi Pico; FIDO2; WebAuthn; CTAPHID; Group Policy; Device Installation Restrictions; endpoint security; Windows.

I. INTRODUCTION

The Universal Serial Bus (USB) trust model was designed for usability rather than security. When a device declares itself a member of the HID class, the host loads a class driver and begins accepting input without prompting for authorization. That decision was reasonable in an era of trusted peripherals. Still, it is the root cause of the BadUSB family of attacks brought to prominence by Nohl and Lell [1], who showed that USB controller firmware can be reprogrammed so that an apparently benign device masquerades as a keyboard. Keystroke-injection attacks weaponize this trust directly. The malicious device carries no executable for an antivirus engine to scan; it merely "types." From the host's perspective, the input is indistinguishable from that of a human operator, which is precisely why the technique bypasses file-based detection and most data-loss-prevention tooling [2]. Once gated behind purpose-built commercial hardware, the capability has been democratized to the point of triviality: a Raspberry Pi Pico costing a few dollars can be flashed with freely available CircuitPython firmware to become a fully functional injection platform [3]. The canonical host-based defense on Windows is to constrain device installation for keyboard/HID setup classes via Group Policy [4]. Practitioner guidance correctly notes that mass-storage controls do not help—a keyboard is not a storage device—and that the HID class cannot be denied wholesale, because the user's legitimate keyboard and mouse are themselves HID devices. This paper makes a further observation that, to our knowledge, has not been systematically analyzed in the academic literature, and which we take as our central contribution: *the same HID-class restriction that defends against keystroke injection can break FIDO2 hardware authentication.* FIDO2/WebAuthn security keys—the strongest widely available defense against credential phishing—

communicate via the CTAPHID protocol, which, by design, rides the USB HID transport for driverless installation [5]. A Device Installation Restriction is strict enough to stop a Pico, therefore, risks turning off the very authentication mechanism the organization deployed against phishing. The defender is placed in a security-versus-security dilemma: mitigating one high-impact attack class can silently weaken defense against another. **Contribution and scope.** This is a *conceptual and policy-analysis* paper with an empirical validation component; it is not a large-scale measurement study. Concretely, we (i) identify and characterise the FIDO2 collateral-damage problem as a first-class security-versus-security conflict; (ii) explain why it arises from the shared HID transport and how the dedicated FIDO usage page (0xF1D0) makes a precise resolution possible; (iii) give a layered, descriptor-aware Windows policy and validate it experimentally on Windows 11 (24H2) with a Pico and two vendors' FIDO2 keys, confirming that the injection defence and phishing-resistant authentication can hold simultaneously; and (iv) analyse the residual identifier-spoofing risk and argue for attestation-based authorisation. Our scope is USB-connected authenticators on Windows; NFC and Bluetooth (caBLE) transports fall outside the HID-blocking dilemma and are not considered. The remainder of the paper is organized as follows. Section II reviews related work. Section III describes the Pico attack platform. Section IV states our analysis method, threat model, and assumptions. Section V sets out the standard Windows defense and its HID-blocking dilemma. Section VI—the core of the paper—formalizes the FIDO2 collateral-damage problem. Section VII proposes a reconciliation and its validation. Section VIII discusses limitations, including identifier spoofing. Section IX outlines future work, and Section X concludes.

II. RELATED WORK

Tian et al. [6] provide the canonical systematization of USB security and situate HID injection within the broader landscape of bus-level threats. Their system contributions are directly relevant to mitigation: GoodUSB [7] mediates device enumeration by reconciling a device's claimed class with user expectations, and USBFILTER [8] implements an in-kernel, packet-level firewall for USB traffic. Gateway-isolation designs such as Cinch [9] interpose a trusted intermediary that can strip unexpected HID traffic from devices that would otherwise be permitted. On detection, Karantzas [2] infers BadUSB activity from native Windows logs by correlating Plug-and-Play device-arrival events with anomalously fast inter-keystroke timing and the proximate spawning of interpreters such as PowerShell. Hardware-assisted and behavioral systems, such as USB-Watch [10] and USBESAFE [11], classify timing and behavioral features in some cases before keystrokes reach the operating system. Pico-specific work is recent and largely applied: Reguly [3] documented turning the Pico into an HID; subsequent projects extend it with wireless command-and-control [12], and the same hardware has even been repurposed as a defensive descriptor inspector [13].

Across this body of work, mitigation is consistently framed as a trade-off between security and usability (the risk of locking out legitimate keyboard users). The trade-off between security and another security control the breakage of FIDO2 authentication by HID-class restrictions is, to our knowledge, not treated as a first-class problem. That gap motivates this paper.

III. THE RASPBERRY PI PICO AS AN ATTACK PLATFORM

The Raspberry Pi Pico is built on the RP2040 microcontroller and exposes a native USB interface. Flashed with CircuitPython and the Adafruit HID library, it emulates a USB keyboard using a few lines of Python placed in a code.py file that the board executes automatically on power-up [3]. No soldering, custom hardware, or specialist toolchain is required, and a minimal payload—opening the Run dialogue, launching an interpreter, and executing a remote script—completes in seconds. The listing in Figure 1 is limited to what is necessary to understand the threat model; it is widely published and does not constitute a novel offensive capability.

```
import time, usb_hid
from adafruit_hid.keyboard import Keyboard
from adafruit_hid.keycode import Keycode
kbd = Keyboard(usb_hid.devices)
time.sleep(2)   # wait for host enumeration
kbd.send(Keycode.GUI, Keycode.R)  # open Run
time.sleep(0.5)
# ... type an interpreter invocation, ENTER
```

Figure 1. Illustrative HID keystroke-injection payload (abridged).

Because no file is written to disk during the keystroke phase, signature-based antivirus provides little protection, and the attack inherits the logged-on user's privilege level [2]. A representative Windows attack chain comprises: (i) delivery (a dropped device or brief unattended access); (ii) enumeration as an HID keyboard, loading the in-box driver; (iii) injection of commands; (iv) fileless execution in memory; and (v) optional cleanup of traces.

IV. ANALYSIS METHOD, THREAT MODEL, AND ASSUMPTIONS

Method. Our analysis proceeds by descriptor-level reasoning about how Windows Device Installation Restrictions interact with the USB HID class, followed by an empirical validation on Windows 11 (Section VII-B) that exercises the four decisive cases on a live host. We reason at the level of setup classes, hardware/instance IDs, and HID usage pages, because these are the objects Group Policy actually matches on.

Threat model. We assume an attacker with brief physical access to a target USB port (a dropped or swapped device, or short unattended access) who can program a Pico to present arbitrary HID descriptors and USB identifiers, but without prior administrative credentials on the host. The defender is a Windows administrator who centrally manages device installations via Group Policy and issues FIDO2 security keys for phishing-resistant authentication. The asset to protect is twofold: the host must reject injection-capable keyboards, and legitimate FIDO2 authenticators must continue to enroll and authenticate.

Assumptions. We assume (a) authenticators expose CTAPHID on the FIDO usage page 0xF1D0, distinct from the generic keyboard usage; (b) the host applies Device Installation Restriction policy at enumeration time and logs the outcome via Kernel-PnP / DeviceSetupManager events [14]; and (c) the administrator can enumerate authorized hardware IDs on a representative machine. Where an authenticator additionally exposes a CCID interface, or where per-OS HID behavior differs, we note the effect in Section VIII rather than assuming uniform behavior.

V. THE STANDARD WINDOWS DEFENCE AND THE HID-BLOCKING DILEMMA

Three commonly deployed controls fail against HID injection, and the reasons matter for what follows. A signature-based antivirus is bypassed because the vector is keystrokes rather than a stored binary. The Removable Storage Access policy denies only the mass-storage class, explicitly leaving keyboards and mice untouched [4]. The legacy technique of disabling the USBSTOR driver likewise does nothing to a device that enumerates as a keyboard.



The native control that can address HID injection is Device Installation Restrictions, configured under

Computer Configuration > Administrative Templates > System > Device Installation > Device Installation Restrictions and persisted under

HKLM\SOFTWARE\Policies\Microsoft\Windows\DeviceInstall\Restrictions [4], [15].

Here lies the dilemma. Mass storage can be denied wholesale because a workstation rarely needs arbitrary USB storage. HID cannot, because the user's own keyboard and mouse belong to the very class one would block. The defender's options are:

- *Blanket class denial* of the keyboard/HID setup classes—effective against the Pico, but it turns off legitimate input devices and, as Section VI shows, FIDO2 keys.
- *Deny-unspecified*, permitting only devices explicitly described by other policy settings—convenient, but it silently captures any HID device not on the allowlist, including authenticators.
- *Hardware-identifier allowlisting*, denying the class but permitting specific Vendor/Product or device-instance IDs—the recommended model, but operationally burdensome and, as Section VIII argues, defeasible by Pico-side spoofing.

The literature treats this dilemma purely as a security-versus-usability problem. The next section shows it is also a security-versus-security problem.

VI. THE FIDO2 COLLATERAL-DAMAGE PROBLEM

1. Why FIDO2 keys ride the HID transport

FIDO2/WebAuthn authentication on a hardware security key is carried out via the Client-to-Authenticator Protocol (CTAP) [5], exposed to the relying party through the WebAuthn API [16]. For USB-connected authenticators, CTAP messages are framed using the HID protocol—a binding the FIDO Alliance calls CTAPHID—so that authenticators install without bespoke drivers on all major host platforms [5]. A USB security key, therefore, presents to Windows as an HID device; driverless installation over HID is a stated design objective of the protocol, not an implementation accident.

Crucially, the authenticator does not use the keyboard HID usage page. CTAPHID is allocated a dedicated FIDO Alliance usage page (0xF1D0) with its own raw input/output reports, distinct from the generic desktop keyboard usage [5]. This distinction is the technical hinge of our argument: it is what makes a precisely scoped defense possible, and what makes a blunt one harmful.

2. The conflict

Consider the three defender options from Section V in light of the above:

- A blanket denial of the HID setup class matches FIDO2 keys because they enumerate within that class. The policy that stops the Pico also stops the security key.

- A deny-unspecified policy denies any device not on the allowlist; unless every authenticator model is enumerated in advance, security keys are blocked on first insertion—precisely when a user tries to register or roam to a new machine.
- Even hardware-identifier allowlisting silently excludes authenticators unless their identifiers are added; an organization that allows only keyboards and mice will find that newly issued or replacement security keys fail to install.

The result is a genuine security-versus-security trade-off, summarised in Table 1. The defender hardening against one of the most effective endpoint-compromise techniques (keystroke injection) can, under the same policy, degrade defense against another prevalent technique (credential phishing), for which FIDO2 is the strongest available countermeasure. The failure mode is especially pernicious because it is silent and delayed: the injection defense appears to work, while the authentication breakage surfaces only when a user later tries to enroll or use a key, often on a machine other than the one where the key was first registered.

Table 1. The security-versus-security trade-off introduced by HID-class restrictions. Only a policy that explicitly accounts for the FIDO transport satisfies both goals.

Policy choice	Stops Pico	Breaks FIDO2 keys
Removable Storage Access only	No	No
Blanket HID class denial	Yes	Yes
Deny-unspecified	Yes	Yes (unless pre-enumerated)
HW-ID allowlist (kbd/mouse only)	Yes	Yes (keys not on list)
HW-ID allowlist incl. keys (§VII)	**Yes**	**No**

VII. RECONCILING INJECTION DEFENCE WITH FIDO2

1. A layered, descriptor-aware policy

The dedicated FIDO usage page (0xF1D0) is what makes reconciliation feasible: a keyboard-emulating attack device and a FIDO2 authenticator are both HID devices, but they are distinguishable at the descriptor level. We recommend the following layered policy.

- *Inventory all authorized HID hardware, including security keys.* On a representative machine, record the hardware IDs and device-instance IDs of every sanctioned keyboard, mouse, and FIDO2 authenticator from Device Manager or the System event log. Desktop peripherals nest several levels deep in the PnP tree; capture each intermediate hub/composite ID.
- *Allow the known-good devices.* Enable installation of devices matching these IDs and add the collected identifiers, treating authenticators as first-class entries rather than an afterthought.
- *Prevent the rest, narrowly.* Prefer denying the keyboard setup class over the entire HID class, and avoid the blunt deny-unspecified switch on any workstation where security keys are used. The objective is to block injection-capable keyboards while leaving the FIDO transport reachable.
- *Preserve administrative recovery.* Enable "Allow administrators to override Device Installation Restriction policies" so a replacement keyboard or key can be installed for legitimate maintenance; scope override rights to a named administrative group, log each override, and pair it with a documented break-glass procedure so recovery does not itself become an injection vector.

- *Validate both properties.* After gpupdate /force, confirm with RSoP/gpresult and the device-installation audit events that an unknown keyboard is rejected, and that an issued security key still enrolls and authenticates. Authentication breakage must be tested explicitly; it does not surface during normal device inventory.

This refinement is the only row in Table 1 that satisfies both security goals—at the cost of an accurate, maintained allowlist that, as Section VIII shows, rests on an assumption the attacker can violate.

2. Validation procedure and results

To move the claim from assertion to demonstration, we implemented the layered policy on a single Windows host and exercised the four decisive cases. The environment was one workstation running Windows 11 (version 24H2) with the descriptor-aware allowlist of Section VII-A applied via local Group Policy; a Raspberry Pi Pico (RP2040) flashed per Figure 1 as an unauthorised keyboard; one authorised USB keyboard; and two FIDO2 security keys from different vendors (a Token2 key and a YubiKey), used to confirm the behaviour is not specific to one authenticator model. Device-installation auditing was enabled so that allow/deny decisions were recorded via Kernel-PnP / DeviceSetupManager events 6416 (install permitted) and 6420 (install blocked by policy) [14]. A policy is correct only if C1 is blocked while C2–C4 succeed simultaneously under the same enforced configuration.

Table 2. Observed results of the four-case validation on Windows 11 24H2 under the descriptor-aware allowlist. All four cases passed: C1 was blocked while C2–C4 succeeded under the same enforced policy.

Case	Device/action under enforcement	Observed outcome	Evidence	Result
C1	Unknown Pico (RP2040) enumerating as keyboard; VID/PID not on allowlist	Install blocked	Event 6420 / RSoP	Pass
C2	Authorized keyboard; VID/PID on allowlist	Install allowed	Event 6416	Pass
C3	FIDO2 key registration / first insertion (0xF1D0 usage)	Enrolment succeeded	WebAuthn ceremony	Pass
C4	FIDO2 authentication after policy enforcement	Auth succeeded	Sign-in / gpresult	Pass

All four cases produced the intended outcome. Cases C1 and C2 confirmed that the injection defense still discriminates between authorized and unauthorized keyboards: the unknown Pico was rejected during enumeration (logged as event 6420), while the allowlisted keyboard installed normally (event 6416). Cases C3 and C4 confirmed that the FIDO transport remained reachable under the same policy: both the Token2 and the YubiKey completed WebAuthn registration on first insertion and subsequently authenticated after policy enforcement. For contrast, we also confirmed that a blanket HID-class denial and a deny-unspecified policy each blocked C3–C4 (both authenticators failed to install), reproducing the collateral-damage effect and matching the corresponding rows of Table 1. Because C3 exercised first insertion on a machine where the key had not previously been registered, it captures the roaming case that most often surfaces the silent breakage in production. These results are illustrative rather than a large-scale study—a single host and two authenticator models—but they demonstrate that the conflict is real and that the descriptor-aware allowlist resolves it in practice, not only in principle.

VIII. LIMITATIONS AND RESIDUAL RISK

Identifier spoofing. Hardware-identifier allowlisting assumes an attacker cannot trivially clone an authorized device's Vendor and Product IDs. That assumption does not hold for the Pico, which can be programmed to present an arbitrary VID/PID and mimic the full descriptor set of a sanctioned device. An attacker who learns the identifiers on an allowlist—or who impersonates a common keyboard model—can craft a rogue device that the policy will admit. The allowlist, therefore, raises the attacker's reconnaissance cost rather than providing an absolute barrier. This residual risk is not a footnote to the design; it is the reason the design is necessary-but-insufficient, and it should be managed with compensating controls: (i) FIDO device attestation, so that issued keys are verified against the FIDO Metadata Service rather than trusted on self-asserted IDs [17]; (ii) behavioural detection of injection, correlating device-arrival events with anomalous inter-keystroke timing and interpreter spawning [2]; and (iii) endpoint monitoring that alerts on new HID enumeration and on policy-override events, so that a spoofed-ID admission is at least visible after the fact.

Maintenance burden and silent breakage. Maintaining an accurate allowlist covering every keyboard, mouse, and authenticator across a heterogeneous fleet—with hot-desking, replacements, and roaming users—imposes real administrative cost, and the FIDO breakage mode remains silent until a user attempts to authenticate. The validation of Section VII-B should therefore be treated as a recurring control rather than a one-time check.

Transport and platform variation. Some authenticators also expose a Smart Card (CCID) interface, and FIDO behavior over HID varies across operating systems [18]; on Windows, direct HID access to a key may also require elevated privileges. Exact policy behavior should be validated for each authenticator model and Windows build. Scope. Our analysis addresses USB-connected authenticators; NFC and Bluetooth (caBLE) transports fall outside the HID-blocking dilemma.

IX. FUTURE WORK

The principal direction follows from the spoofing limitation: because the Pico can forge hardware identifiers, authorization must not rest on self-asserted, spoofable identifiers. Promising avenues include (i) cryptographic device attestation, in which a peripheral proves possession of a hardware-bound key before the host accepts its input; (ii) behavioural binding that augments descriptor checks with continuous keystroke-dynamics verification, so that a device presenting valid identifiers must also exhibit human-consistent timing; and (iii) bus-level mediation extending designs such as Cinch [9] that authorise input out of band rather than trusting a device's claims. An empirical study quantifying how reliably a spoofed-identifier Pico defeats a fully configured Windows allowlist—and how much reconnaissance this requires—would itself be a valuable contribution. A complementary strand is to propose OS or policy-template mechanisms that block injection-capable keyboards while reliably exempting the CTAPHID usage page, so that the conflict identified here is resolved by design rather than by careful manual configuration.

X. CONCLUSION

The Raspberry Pi Pico exemplifies how a once-specialized attack has become trivially accessible, and the natural Windows defense—restricting HID device installation—appears straightforward. This paper has shown that the defense carries a hidden cost that the existing security-versus-usability literature has overlooked: a HID-class restriction strict enough to prevent keystroke injection can simultaneously break FIDO2 hardware authentication, because security keys ride the same HID transport as keyboards. Hardening against one high-impact attack can thus weaken defense against another. The conflict is resolvable—the FIDO usage page is distinct from the keyboard usage, so a precisely scoped allowlist that enumerates authenticators can satisfy both goals, as we demonstrated on Windows 11 (24H2), where a Pico was blocked while two vendors' FIDO2 keys

still enrolled and authenticated under the same policy—but only if defenders recognise the problem, and only as far as spoofable identifiers allow. Recognizing the FIDO2 collateral-damage problem and ultimately replacing identifier-based authorization with attestation are necessary if endpoints are to resist keystroke injection without sacrificing phishing-resistant authentication.

REFERENCES

[1]. K. Nohl and J. Lell, "BadUSB—On accessories that turn evil," presented at Black Hat USA, Las Vegas, NV, USA, 2014.
[2]. G. Karantzas, "Forensic log-based detection for keystroke injection 'BadUSB' attacks," arXiv:2302.04541, 2023.
[3]. T. Reguly, "Turning a Pico into a Human Interface Device (HID)," Tripwire State of Security, 2022.
[4]. Microsoft, "Manage device installation with Group Policy," Microsoft Learn, 2025.
[5]. FIDO Alliance, "Client to Authenticator Protocol (CTAP) 2.1," Proposed Standard, 2021.
[6]. J. Tian, N. Scaife, D. Kumar, M. Bailey, A. Bates, and K. Butler, "SoK: 'Plug & Pray' today—Understanding USB insecurity in versions 1 through C," in Proc. IEEE S&P, 2018, pp. 1032–1047.
[7]. D. J. Tian, A. Bates, and K. Butler, "Defending against malicious USB firmware with GoodUSB," in Proc. 31st ACSAC, 2015, pp. 261–270.
[8]. D. J. Tian, A. Bates, and K. Butler, "Making USB great again with USBFILTER," in Proc. 25th USENIX Security, 2016, pp. 415–430.
[9]. S. Angel et al., "Defending against malicious peripherals with Cinch," in Proc. 25th USENIX Security, 2016, pp. 397–414.
[10]. K. Daley, K. Morris, P. Mensah, and A. Bates, "USB-Watch: A dynamic hardware-assisted USB threat detection framework," in Proc. 15th SecureComm, 2019, pp. 126–146.
[11]. A. Kharraz, K. Daley, G. Baker, W. Robertson, and E. Kirda, "USBESAFE: An end-point solution to protect against USB-based attacks," in Proc. 22nd RAID, 2019, pp. 89–103.
[12]. A. Sebastian and R. Kumar, "Wireless Bad USB using Raspberry Pi Pico W," Int. J. Emerging Technologies and Innovative Research, vol. 11, no. 4, pp. c372–c375, 2024.
[13]. V. Suresh and A. Nair, "USB HID intrusion detection and control system: A BadUSB detector using Raspberry Pi Pico," Int. J. Innovative Research in Science, Engineering and Technology, vol. 14, no. 5, 2025.
[14]. Microsoft, "Device installation event log reference (DeviceSetupManager / Kernel-PnP)," Microsoft Learn, 2024.
[15]. W. Gabel, "Restrict installation of (USB) devices using Group Policy," 4sysops, 2023.
[16]. W3C, "Web Authentication (WebAuthn) Level 2," W3C Recommendation, 2021.
[17]. FIDO Alliance, "FIDO Metadata Service (MDS) v3.0 and authenticator attestation," 2021.
[18]. Token2, "Understanding FIDO2 authentication across different operating systems and browsers," 2024.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

AI in Decision-Making: The Case of University Metropolitan Tirana

Suada Gjimi
Department of Economic Informatics
University Metropolitan Tirana
Tirana, Albania
suada.peci@umt.edu.al
ORCID: 0009-0003-1226-0387

Arben Asllani
Department of Data Analytics
University of Tennessee
Chattanooga, United States
beni-asllani@utc.edu
ORCID: 0000-0001-5425-6763

Abstract
The integration of artificial intelligence into higher education has fundamentally changed the way it operates. Employees increasingly limit critical thinking by delegating their work to AI systems. The use of AI tools raises considerable concerns, especially in decision-making processes. Recent empirical studies examine academic staff, but lack a parallel with administrative staff. They examine the spectrum from human-led augmentation to AI-driven delegation. This study presents an exploratory survey conducted at the University Metropolitan Tirana (UMT), an early-adopter institution in integrating AI into the workplace. The survey examines 57 employees (response rate: 63.3%) who use AI tools in their professional decision-making. The findings indicate a high level of AI use (94.7%), with daily use of AI tools resulting in 66.7%. Academic (M = 3.23) and administrative (M = 3.14) decisions are surpassed by strategic decisions (M = 3.51) when it comes to relying on AI recommendations. This pattern overturns the assumption that "routine tasks come first," prevalent in the existing literature. Role-based differences show that administrative staff exhibits a considerably higher concentration of delegation behavior (25% Level 3) compared to academic staff (2.7%). The study contributes the first survey-based investigation of AI-driven decision-making among university employees in Albania, the ADIS framework as a measurable taxonomy of the augmentation–delegation spectrum, and a structured survey instrument for application across regional institutions.

Keywords: generative AI, decision-making, higher education, AI adoption

I. INTRODUCTION

Generative Artificial Intelligence (GenAI) is part of organizational decision-making, reshaping how employees approach intensive tasks. Replacing old decision-making methods where employees had to access information, analyze, and make decisions. Recent surveys and studies highlight the use of GenAI in task execution, increasing quality and efficiency, transforming coordination structures, and enhancing managerial control in decision-making processes [1], [2], [3]. Organizational decisions made by AI systems carry implications regarding decision quality and accountability. Since AI is not transparent in how it provides recommendations, it creates an accountability gap among the system, the individual, and the organization [4]. Universities provide a perfect environment for examining these dynamics because their decisions span academic, administrative, and strategic domains. This makes the risks of misdelegation diverse and, in many cases, irreversible. The focus of this study is the case of the University Metropolitan Tirana, an early adopter of AI, where ongoing training is provided to help staff use AI tools in their daily work. The university has actively integrated AI tools, making it a natural laboratory for studying AI use in decision-making. Staff training ensures efficiency and faster decision-making. There are several studies on AI in higher education, but their focus is on student adoption, pedagogical applications, and academic integrity. These studies do not highlight how employees make decisions in an AI-integrated work environment [5], [6]. This paper addresses this gap through an exploratory survey of employees at the University of Metropolitan Tirana. This paper addresses the gap through the following research questions:

RQ1: What AI tools are employees at UMT currently using in their work?

RQ2: For which types of decisions do employees use AI, and to what degree?
RQ3: Do employees perceive themselves as using AI for augmentation or delegation?
RQ4: Are there significant differences in AI use behavior by employee role?

This study contributes: (a) empirically, the first survey-based investigation into AI-driven decision-making behavior among university employees in Albania; (b) theoretically, it proposes the AI Decision Integration Spectrum (ADIS), a three-level framework that classifies employees' decision-making behavior as AI-assisted, AI-recommended, or AI-delegated; (c) practically, it provides a replicable survey instrument and an interpretive framework that regional universities can apply as they navigate AI adoption across their own academic, administrative, and strategic functions.

II. RELATED WORK

1. Information-Seeking to AI-Mediated Decision-Making

Information search and organizational decision-making were inseparable activities before the use of AI. At the center is the human who seeks a specific item by interpreting the decision-making path [7]. Information search begins with the filling of a knowledge gap in the individual, who actively works to fill it through independent discovery, evaluation, and synthesis. The integration of Artificial Intelligence has disrupted the architecture. The role of the individual has shifted from synthesizer to evaluator, taking advantage of AI's rapid data processing. Attitudes towards AI-driven decision-making are shaped by individual efficiency and benefits, according to empirical studies to date [8]. Existing literature defines automation as a complement to, rather than a substitute for, human skills. Machines perform human tasks faster, but human-machine collaboration is increasing [4]. The literature highlights a paradox: in practice, the boundary between substitution and augmentation is blurred. The path to automation, starting with organizations, ends with workers accepting AI's results without further criticism. As a result, concerns about accountability, fairness, and transparency of systems continue [9]. Automation bias is defined as the tendency of humans to trust the automated recommendations, even when inaccurate information is noticed [10], [11]. Studies confirm that automation bias reduces sensitivity but cannot eliminate it [12]. Delegation is not determined by individual failure. The findings suggest that organizations should actively train staff to minimize the risk of delegating decisions. This paper examines, through the framework of the AI Decision Integration Spectrum, the extent to which this tendency manifests itself across roles in employees of the University of Metropolitan Tirana.

2. AI Adoption in Higher Education

The integration of artificial intelligence in higher education has become an indispensable tool in resource allocation, information processing, data analysis, and institutional decision-making. This is confirmed by numerous studies in this field, which, in addition to positive effects, have found significant dependencies among the challenges of GenAI adoption. Most of the uncertainties concern the limits of use and the strategic frameworks needed to keep pace with adoption rates [13], [14], [15]. Existing studies focus more on student behavior and the impact of AI, overshadowing research on staff within universities. Currently, studies on staff have been more complex, where the use of AI is guided by ethical considerations, suggesting that faculty use of AI should be shaped by professional values [16]. Empirical findings highlight that integrating AI into the workplace entails different engagement models. ChatGPT is the most widely used tool among academic staff, adapted to their needs. It reduces stress levels and increases positive affect among faculty [17]. On the other hand, the administrative staff requires further research. So far, it has emphasized the use of AI tools in the design of institutional communications, data organization, and related areas, and has positively evaluated the integration of AI. Although recent findings consider administrative staff to be capable adopters of AI, there is a lack of studies on their behavior, with most focusing on academic staff and students [18].

3. Decision-Delegation Problem

The integration of AI into organizational life has enabled a shift in one of the most important tasks from automation to delegation. Task automation replaces procedural activities such as planning, data collection, and document processing. Decision - delegation is different from automation, it transfers judgment to an AI system while accepting responsibility for outcomes that affect people. While large language models produce and imitate expert reviews with high efficiency, users avoid prior control. A large-scale empirical survey of knowledge workers has found that workers' critical-thinking effort decreases when generative AI tools appear to be efficient at completing a task [19]. A review of human-AI augmentation in the workplace finds that the spectrum between augmentation and delegation lacks formal measurement frameworks at the organizational level. Institutional and organizational factors moderate the extent to which AI functions as a cognitive tool or a cognitive substitute [20]. In cases where AI systems take over decision-making functions, accountability issues become an intractable concern, as defined in the literature as a failure in the governance and contemporary deployment of AI [21].

III. METHODOLOGY

This study used an exploratory survey to study the impact of AI on decision-making among staff at the University of Metropolitan Tirana. The survey was sent to 90 university employees, including academic and administrative staff. For studying the demographics, use of AI tools, decision-making behavior, perceptions, and trust, as well as an open-ended question on concerns about using AI at work. The survey included 16 questions organized into sections and presented in various formats, including multiple-choice, five-point Likert, and open-ended. UMT employees completed the questionnaire anonymously through Microsoft Forms. For data analysis, Python is used to report frequencies, percentages, means, medians, and standard deviations. Open-ended responses were analyzed thematically using inductive coding. Participation was voluntary and anonymous, with no personally identifiable data collected. The survey is approved by the University of Metropolitan Tirana and conducted in accordance with established ethical standards for research involving human subjects. Participation was voluntary and anonymous, with no personally identifying data collected; participants were informed of the study's purpose, the voluntary nature of participation, and the confidential handling of their responses through the consent statement at the beginning of the Microsoft Form.

IV. RESULTS

A total of 57 valid responses were received, with a response rate of 63.3%. The majority of responses were from academic staff, with 42 responses (73.3% response rate). At the same time, the administrative and management staff yielded 15 responses (26.3%).

RQ1: What AI tools are employees at UMT currently using in their work?
Survey results show that employees have a high level of engagement with AI tools. As shown in Figure 1, ChatGPT was used by 87.7% of respondents, making it the most widely used tool. The following tools are Google Gemini (47.4%) and Claude (35.1%), while Microsoft Copilot (10.5%) had a lower engagement rate.

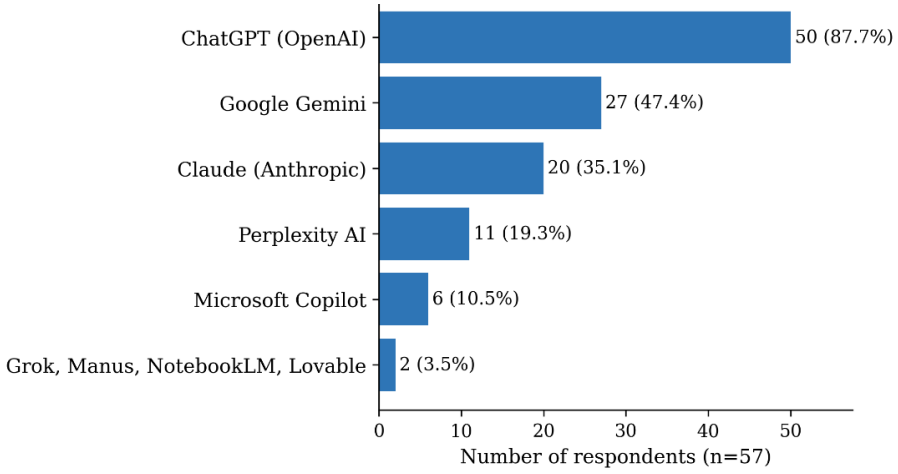


Figure 1. AI tools are currently used by employees at University Metropolitan Tirana (UMT).

Although use was intensive, only three respondents reported not using AI tools. 38 (66.7%) respondents report using AI tools daily, 12 (21.1%) weekly, and 4 (7%) rarely. Considering the profile of the respondents, academic staff show the highest rate of daily use of artificial intelligence tools (69%), followed by administrative staff (60%).

Table 1 summarizes the results of the task profile on which employees most often use AI tools, with multiple-select options. Researching and finding information is the most common task (73.7%), followed by data analysis and summarizing (59.6%), supporting decision-making (57.9%), and drafting emails or documents (56.1%). Notably, more than half of respondents reported explicit use of AI for "supporting decisions or recommendations," a task category directly relevant to RQ2 and RQ3.

Table 1. Task Types for which AI is Used.

Task	***n***	**% of respondents**
Researching	42	73.7%
Data analysis or summarizing	34	59.6%
Supporting decisions	33	57.9%
Drafting emails or documents	32	56.1%
Generating reports	18	31.6%
Student or colleague advising	17	29.8%
Scheduling or planning	7	12.3%
Coding	2	3.5%

Tasks differ by role, and academic staff reported using ChatGPT (88.1%), Google Gemini (57.1%), Claude (42.9%), and Perplexity AI (26.2%). The administrative staff focused primarily on ChatGPT (86.7%), with lower

adoption of Gemini (20%) and Claude (13.3%). This pattern suggests that academic staff treat AI primarily as a research tool, while administrative staff use it more transactionally, focusing on it as a tool.

RQ2: For which types of decisions do employees use AI, and to what degree?
Respondents reported on AI recommendations in three different decision categories. Strategic decisions had the highest mean reliance score (M = 3.51, SD = 1.48), followed by academic decisions (M = 3.23, SD = 1.52), and administrative decisions had the lowest score (M = 3.14, SD = 1.39). Given the importance of decision-making, which is often considered the least automatable category, survey reports show that 63.2% of employees rely on AI for strategic decisions (Table 2a).

In the behavioral analysis, respondents are divided into two groups: 22 (38.6%) say they always or usually verify AI results with other sources. In contrast, 26 (45.6%) report trusting AI and never or rarely checking the answers received (Table 2b). Considering the above decision-making results and employees' behavior towards AI tools, a high number of employees act according to AI recommendations without independent verification, a behavioral pattern directly related to the augmentation-delegation distinction examined in RQ3.

Table 2. AI reliance and verification behavior.

(a) Reliance on AI by Decision Type (1 = Not at all, 5 = Completely).

Decision Type	***M***	***Mdn***	***SD***	**% Often / Compl.**	**Mann-Whitney U**	***p***	***r***
Administrative	3.14	3	1.39	42.1%	395.0	.141	-0.25
Academic	3.23	4	1.52	52.6%	533.0	< .001	-0.69
Strategic	3.51	4	1.48	63.2%	449.0	.012	-0.43

(b) Verification of AI Recommendations.

Frequency	***n***	**%**
Never	15	26.3%
Rarely	11	19.3%
Sometimes	9	15.8%
Usually	10	17.5%
Always	12	21.1%

RQ3: Do employees perceive themselves as using AI for augmentation or delegation?

The AI Decision Integration Spectrum (ADIS) is a three-tier framework developed after responders were classified into three categories based on their self-reported decision-making behavior. As shown in Figure 2, the majority of employees fell into Tier 2 (AI-Recommended), with 26 responders representing 53.1% of employees. While 19 employees (38.8%) reported that they make the decision themselves after gathering information, falling in Tier 1 (AI-Assisted), and 4 employees (8.2%) reported full delegation behavior, consistent with Tier 3 (AI-Delegated). The remaining 8 respondents (14%) who selected “I do not use AI for decisions” were excluded from the tier analysis.

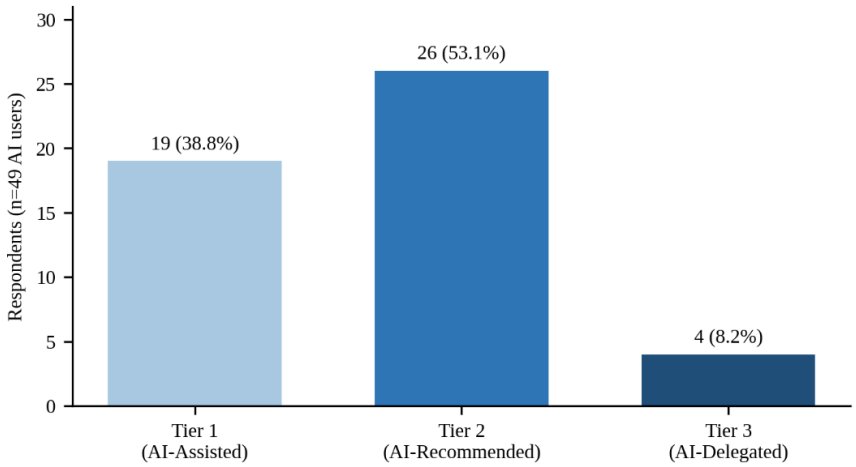


Figure 2: ADIS tier distribution among AI users (n = 49).

Examining mean scores within each ADIS tier reveals a clear gradient supporting the framework's theoretical structure. As shown in Table 3, trust in AI recommendations rises monotonically from Tier 1 (M = 3.37) to Tier 2 (M = 3.73) to Tier 3 (M = 4), while verification behavior falls in the opposite direction: 47.4% of Tier 1 respondents always or usually verify AI outputs, compared with 26.9% of Tier 2 respondents and 0% of Tier 3 respondents. The augmentation-delegation paradox, highlighted in the existing literature, concerns the relationship employees form with AI systems: greater trust and less verification. Employees are moving from AI-assisted work to direct delegation.

Table 3. Within-Tier Means on Key Indicators.

Indicator	**Tier 1 (n=19)**		**Tier 2 (n=26)**		**Tier 3 (n=4)**		**Spearman ρ**	***p***
	M	***SD***	***M***	***SD***	***M***	***SD***		
Trust in AI recommendations	3.37	0.76	3.73	0.87	4.00	0.00	+0.27	.057
Confidence in evaluating AI output	3.95	0.97	4.23	0.71	4.00	0.82	+0.08	.575
AI improved decision quality	3.79	1.18	4.31	0.68	4.50	0.58	+0.23	.112
% verifying Always/Usually	47.4%	-	26.9%	-	0.0%	-	-0.33	.021

With only 4 respondents in Tier 3, this group alone cannot support strong conclusions. The values, however, follow the same direction as Tiers 1 and 2, providing initial support for the view that the three-tier classification reflects a meaningful progression. Tier 2 represents the modal pattern of AI use at UMT: high trust (3.73); strong confidence in evaluating AI output (4.23); and the highest perceived quality improvement (4.31), paired with lower verification behavior than Tier 1.

Empirical support is provided for the augmentation-delegation paradox at the individual level. Trust in AI recommendations and frequency of verification are significantly correlated (Spearman $\rho = -0.36$, $p = .011$, n = 49, medium effect size).

RQ4: Are there significant differences in AI use behavior by employee role?
Table 4 shows the behavioral indicators of academic and administrative staff. Academic staff are represented by higher values regarding reliance on AI across all three decision types. It also results in stronger trust in AI and in the evaluation of its results. The role is decisive in the levels of ADIS. The rate of AI use among academic staff is 37.8% at Level 1 and only 2.7% at Level 3. According to the survey report, 25% of administrative staff were placed at Level 3. However, administrative staff had a higher frequency of result verification, at 46.7%, compared to academic staff, at 35.7%. These patterns are only descriptive; the small administrative subgroup excludes statistical testing.

Table 4. Comparison of Indicators by Role.

Indicator	**Academic (n=42)**		**Administrative (n=15)**		**Mann-Whitney U**	***p***	***r***
	M	***SD***	***M***	***SD***			
Daily AI use	69.0%	-	60.0%	-	-	-	-
Reliance: administrative decisions	3.31	1.30	2.67	1.59	395.0	.141	-0.25
Reliance: academic decisions	3.74	1.29	1.80	1.21	533.0	< .001	-0.69
Reliance: strategic decisions	3.81	1.35	2.67	1.54	449.0	.012	-0.43
Trust in AI recommendations	3.55	0.92	3.20	0.86	383.0	.184	-0.22
Confidence in evaluating AI output	4.07	0.95	3.60	0.91	412.5	.060	-0.31
Perceived quality improvement	4.00	1.08	3.60	1.06	390.0	.155	-0.24
% Tier 1 of users	37.8%	-	41.7%	-	-	-	-
% Tier 2 of users	59.5%	-	33.3%	-	-	-	-
% Tier 3 of users	2.7%	-	25.0%	-	-	-	-
% verifying Always/Usually	35.7%	-	46.7%	-	-	-	-

Employee concerns raised about AI use.
The survey included an optional open-ended question about concerns regarding the use of AI tools in workplace decisions, to which 21 responses were received. The major concerns were related to loss of independent thinking and over-reliance on artificial intelligence. Concerns about accuracy and inconsistency, privacy and data security, reduced creativity, and the costs of AI tools were secondary themes. The qualitative concern with over-dependence aligns directly with the quantitative pattern observed in RQ3: as respondents move along the ADIS spectrum toward delegation, verification behavior weakens. The fact that employees themselves identify this as a concern suggests an awareness gap between recognized risk and reported behavior.

V. DISCUSSION

First, a survey of 57 employees at the University of Metropolitan Tirana found that 94.7% used AI tools. At the same time, 66.7% of employees reported using AI daily in their work. Studies to date have reported on the use of artificial intelligence in higher education, and the findings of this paper are at a high level compared to current reports [13], [15], [22]. Considering that University Metropolitan Tirana positions itself as an AI-focused institution and is active in training staff to use AI tools, the observed rates are high among universities in the region. ChatGPT leads in usage by 87.7% of employees, and other tools that can be considered secondary are Google Gemini with 47.4% and Claude with 35.1%. These findings establish OpenAI's product as the standard among academic and administrative staff [17], [23]. Second, reliance on recommendation systems shows high results with strategic decisions (M = 3.51). While the existing literature reports that delegation begins with routine tasks, this pattern complicates assumptions about strategic decision outcomes. Regarding administrative decisions, the mean reliance score is the lowest (M = 3.14), and the average for academic decisions (M = 3.23) [4], [9]. At the University Metropolitan Tirana, AI is placed at the center of organizational decisions where the consequences are broader. Third, there is the empirical result obtained from the ADIS levels. Trust in AI tools and perceived quality differed across levels, with trust increasing from 3.37 at Level 1 (AI-Assisted) to 4.00 at Level 3 (AI-Delegated). On the same 1-5 scale, the frequency of verification decreased from 3.21 to 1.25. Given that trust and frequency of verification are significantly and inversely correlated, the stratum-level gradient is reflected at the individual level (Spearman $\rho = -0.36$, $p = .011$). The empirical findings in this study support what previous literature has characterized theoretically. The risk that organizations may start with augmentation is that they may slide towards automation. This risk is defined as the augmentation-automation paradox, in which a measurable behavioral gradient is observed at the individual level [4]. The mean verification score of 1.25 at Level 3 demonstrates this tendency in contemporary LLM settings [12]. Finally, concerns and perceived risk related to the use of AI tools. Employees continue to trust the results generated by AI systems, despite being aware of the risks. This confirms that AI tools reduce the effort of critical thinking by encouraging excessive reliance [19]. UMT employees are aware of the risks of delegation regardless of their behavior.

VI. CONCLUSION

The study on 57 employees of University Metropolitan Tirana (63.3% response rate) examined the use of AI in their professional decision-making. Reliance on AI rises with the breadth of the decision, peaking on strategic decisions. The three-tier structure of the proposed AI Decision Integration Spectrum (ADIS) is consistent with the observed patterns in the data: trust increases steadily, while verification declines steadily. The contributions offered in this paper are the first survey-based investigation of AI-driven decision-making among university staff in Albania, the ADIS framework as a measurable taxonomy of the augmentation-delegation spectrum, and a replicable survey instrument for application elsewhere. For universities, the findings suggest that AI policies should match the type of decision to the level of AI use allowed [6]. This matters because at UMT, the decisions that carry the highest risk are also the ones most often delegated to AI. For researchers, the ADIS framework offers a replicable tool. The inverse gradient between trust and verification, and the split among administrative staff, can be tested as hypotheses in future comparative work. This study is limited by its single-institution scope, modest sample size, and the size of the Tier 3 group for strong conclusions. Future work should test whether the UMT models generalize to the majority of institutions in the region. This study suggests that AI in decision-making is best understood not as a tool that staff either adopt or resist, but as a gradual transfer of judgment that institutions must learn to recognize, measure, and govern.

REFERENCES

A. Cvetkovic, I. Savolainen, M. Celuch, M. Heiskari, E. S. Ruokosuo, P. Arriaga, M. Koike, and A. Oksanen, "User trust in AI and major tech companies in twelve countries," Behavior & Information Technology, pp. 1-17, 2026.

[2]. Q. Chen, U. Tawilapakul, and A. M. Lin, "Teacherness in the age of GenAI: An ethnographic exploration of pedagogical decision-making and AI integration in English language teaching," System, vol. 136, p. 103877, 2026.
[3]. M. Fernández-Fernández, Á. Hernández-Tamurejo and P. González-Padilla, "The AI trust factor: an MCA analysis of automation and decision-making in entrepreneurship," International Entrepreneurship and Management Journal, vol. 22, no. 34, 2026.
[4]. S. Raisch and S. Krakowski, "Artificial Intelligence and Management: The Automation–Augmentation Paradox," Academy of Management Review, vol. 46, no. 1, pp. 192-210, 2021.
[5]. O. Gjermeni, "Artificial intelligence perceptions in higher education: A comprehensive analysis in the Albanian context," in Proc. International Conference on New Research in CSIT, 2023.
[6]. Hodaj and S. Laçka, "AI, ethics, and human-centered policy in Albanian education," Interdisciplinary Journal of Research and Development, vol. 12, no. 3, p. 156, 2025.
[7]. G. Marchionini, "Exploratory search: from finding to understanding," Communications of the ACM, vol. 49, no. 4, pp. 41-46, 2006.
[8]. G. Cao, Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Understanding managers' attitudes and behavioral intentions towards using artificial intelligence for organizational decision-making," Technovation, vol. 106, p. 102312, 2021.
[9]. Y. R. Shrestha, S. M. Ben-Menahem, and G. von Krogh, "Organizational decision-making structures in the age of artificial intelligence," California Management Review, vol. 61, no. 4, pp. 66-83, 2019.
[10]. R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Human Factors, vol. 39, no. 2, pp. 230-253, 1997.
[11]. L. J. Skitka, K. L. Mosier, and M. Burdick, "Does automation bias decision-making?" International Journal of Human-Computer Studies, vol. 51, no. 5, pp. 991-1006, 1999.
[12]. G. Romeo and D. Conti, "Exploring automation bias in human--AI collaboration: a review and implications for explainable AI," AI & Society, vol. 41, pp. 259-278, 2026.
[13]. R. Sahar and M. Munawaroh, "Artificial intelligence in higher education with bibliometric and content analysis for future research agenda," Discover Sustainability, vol. 6, p. 401, 2025.
[14]. R. J. Isaifan, "Artificial intelligence in higher education: A global statistical synthesis for policy and quality assurance reform," Education Sciences, vol. 16, no. 3, p. 483, 2026.
[15]. L. Hughes, T. Malik, S. Dettmer, A. S. Al-Busaidi, and Y. K. Dwivedi, "Reimagining higher education: Navigating the challenges of generative AI adoption," Information Systems Frontiers, 2025.
[16]. M. I. Baig and E. Yadegaridehkordi, "Factors influencing academic staff satisfaction and continuous usage of generative artificial intelligence (GenAI) in higher education," International Journal of Educational Technology in Higher Education, vol. 22, no. 5, 2025.
[17]. J. J. Cambra-Fierro, M. Fuentes Blasco, M.-E. López-Pérez and A. Trifu, "ChatGPT adoption and its influence on faculty well-being: An empirical research in higher education," Education and Information Technologies, vol. 30, pp. 1517-1538, 2025.
[18]. S. C. Kong and W. Hu, "A study of developing administrative staff's conceptual understanding of generative artificial intelligence through professional development," Computers and Education: Artificial Intelligence, vol. 9, p. 100444, 2025.
[19]. H.-P. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson, "The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers," in Proc. 2025 CHI Conference on Human Factors in Computing Systems, 2025.
[20]. T. Nguyen and A. Elbanna, "Understanding Human-AI Augmentation in the Workplace: A Review and a Future Research Agenda," Information Systems Frontiers, 2025.

[21]. Y. Zhang, J. Guo, Y. Wang, S. Li, Q. Yang, J. Zhang, and Z. Lu, "Understanding Trust and Willingness to Use GenAI Tools in Higher Education: A SEM-ANN Approach Based on the S-O-R Framework," Systems, vol. 13, no. 10, 2025.

[22]. N. H. Nguyen, N. S. Pham, T. M. Nguyen, T. M. A. Nguyen, V. A. Tran, and T. M. Nguyen, "Assessing the Adoption of Artificial Intelligence in Higher Education: A Case Study of Hanoi Metropolitan University," International Journal of Emerging Technologies in Learning (iJET), vol. 21, no. 1, pp. 73-92, 2026.

[23]. Shata and K. Hartley, "Artificial intelligence and communication technologies in academia: faculty perceptions and the adoption of generative AI," International Journal of Educational Technology in Higher Education, vol. 22, no. 14, 2025.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Emerging Technology Adoption in TVET: A Theoretical and Pedagogical Synthesis

Twaambo Chiinza
Department of Applied Sciences
Technical and Vocational Teachers' College
Luanshya, Zambia
twaambo.chiinza@gmail.com
ORCID:0009-0003-5549-0326

Alice P. Shemi
Information Systems & Technology, School of ICT
Copperbelt University
Kitwe, Zambia
shemiap@gmail.com
ORCID:0000-0003-2951-7743

Abstract
Emerging technologies such as artificial intelligence, learning analytics, simulation, extended reality, and digital platforms are reshaping how Technical and Vocational Education and Training (TVET) systems design learning, support practice, and assess occupational competence. However, their effective use cannot be explained by technology availability or user acceptance alone. This paper presents a structured, literature-based synthesis of theoretical and pedagogical perspectives on the adoption of emerging technologies in competency-based TVET systems. It argues that adoption should be understood as an alignment issue involving technological affordances, pedagogical design, institutional readiness, and competency-based assessment. Using Zambia's Technical Education, Vocational and Entrepreneurship Training context as an illustrative case, the paper contributes a conceptual alignment framework for analyzing technology-mediated vocational education and future empirical validation.

Keywords: emerging technologies, TVET, TEVET, technology adoption, vocational pedagogy, competency-based assessment

I. INTRODUCTION

Technical and Vocational Education and Training (TVET) plays a central role in preparing learners for employment, entrepreneurship, innovation, and lifelong skills development. Its relevance has become more pronounced as digitalization, automation, and data-driven work practices continue to reshape occupational tasks across different sectors. As workplaces increasingly depend on digital systems, artificial intelligence, automation, data analytics, and networked technologies, vocational education systems are expected to prepare learners not only for existing jobs but also for changing work environments where digital competence, adaptability, problem-solving, and continuous learning form part of technical performance itself [1-3]. Emerging technologies have therefore become an important area of research and policy attention in TVET. In this paper, emerging technologies refer to digital tools and systems whose educational use is still developing, but which have the potential to influence how vocational learning is designed, delivered, supported, and assessed. These include artificial intelligence, learning analytics, simulation technologies, extended reality, digital learning platforms, and other interactive or data-driven learning environments. From a computer science and information systems perspective, these technologies are not simply teaching aids. They are computational and socio-technical systems that mediate access to learning, generate data, support decision-making, structure feedback, and influence how evidence of learning is captured and interpreted. When appropriately designed, emerging technologies can support simulated work practice, personalized feedback, flexible access to learning resources, analytics-informed learner support, digital portfolios, and stronger links between classroom learning and workplace expectations [4], [5]. Recent guidance on artificial intelligence in TVET also emphasizes that AI is increasingly influencing teaching, learning, assessment, learner support, and institutional management, while creating new demands for governance, ethical use, staff capacity, and institutional readiness [6]. However, these

possibilities do not automatically translate into improved occupational competence. A digital platform, simulation tool, or AI-supported system may increase access and learner engagement but still fail to strengthen vocational learning if it is weakly aligned with curriculum outcomes, assessment standards, lecturers' capabilities, institutional infrastructure, or industry expectations. This issue is particularly important in competency-based TVET systems. Vocational education differs from general education because its central concern is not only knowledge acquisition, but the development and verification of occupational competence. Competency-based training requires learners to demonstrate what they can do against defined standards, often in contexts that resemble or connect to real work [7], [8]. For that reason, technology-enhanced learning in TVET must be judged by more than the availability, novelty, or technical sophistication of digital tools. Its value depends on whether it supports credible learning, valid assessment, and meaningful progression towards occupational performance. Existing literature provides useful but partial explanations of this problem. Technology adoption theories explain factors such as perceived usefulness, ease of use, facilitating conditions, user acceptance, and institutional readiness [9]. Socio-technical perspectives show that technologies are shaped by organizational routines, infrastructure, leadership, policy, governance, and quality assurance arrangements [10], [11]. Pedagogical theories such as constructivism, situated learning, and communities of practice explain how learners build knowledge through active engagement, authentic practice, social participation, and guided feedback [12-14]. These perspectives are valuable, but they are often applied separately. As a result, technology integration in TVET is frequently discussed either as a matter of adoption, pedagogy, infrastructure, or assessment, rather than as an alignment problem across these interdependent areas. This separation creates a conceptual gap. In competency-based vocational education, the adoption of emerging technologies is not only a technical or behavioral issue. It is also a pedagogical, institutional, and quality assurance issue. A technology must fit the occupational task, the learner's level of competence, the lecturer's pedagogical capability, the available infrastructure, the assessment model, and the policy environment. Where these elements are misaligned, technology integration may remain fragmented, superficial, or difficult to sustain. This is especially important in resource-constrained contexts, where connectivity, equipment, digital skills, maintenance systems, and funding affect whether emerging technologies can be used equitably and effectively [3], [15]. Recent TVET guidance also warns that AI and related technologies can pose risks to infrastructure, costs, sustainability, capability gaps, governance, equity, and assessment integrity if introduced without sufficient institutional preparation [6]. Although this paper uses the broader international term TVET, the Zambian context provides an important illustration of this wider challenge. In Zambia, the sector is formally referred to as Technical Education, Vocational and Entrepreneurship Training (TEVET), reflecting a national policy orientation that integrates technical, vocational, and entrepreneurship training into a single skills development system [16], [17]. Zambia's TEVET system is shaped by competency-based training, institutional accreditation, open, distance, and flexible learning initiatives, quality assurance expectations, and growing policy attention to digital transformation. These conditions make the Zambian case relevant for examining how emerging technologies can be adopted without weakening the credibility of practical training, assessment, and certification. This paper examines the adoption of emerging technologies in TVET from theoretical and pedagogical perspectives. It is guided by the following research question: What theoretical and pedagogical perspectives explain the adoption of emerging technologies in competency-based TVET systems? The paper does not evaluate a single technology or present empirical field findings. Instead, it provides a structured literature-based synthesis to clarify how different theoretical and pedagogical perspectives explain the adoption and use of emerging technologies in vocational education. The paper's contribution is threefold. First, it compares selected theoretical and pedagogical perspectives relevant to the adoption of emerging technologies in TVET. Second, it identifies the limitations of treating technology adoption, pedagogy, institutional readiness, and assessment as separate issues. Third, it proposes an alignment-oriented synthesis to understand the integration of emerging technologies in competency-based TVET systems, using Zambia's TEVET context to illustrate the practical importance of aligning technological affordances, pedagogical design, institutional readiness, and competency-based assessment.

II. REVIEW METHODOLOGY

This paper adopts a structured literature-based review methodology. It is not presented as a systematic review or meta-analysis because its purpose is not to measure the effects of technology or to exhaustively map all publications in the field. Instead, it synthesizes theoretical and pedagogical perspectives to explain the adoption of emerging technologies in competency-based TVET systems. The review was guided by the question: What theoretical and pedagogical perspectives explain the adoption of emerging technologies in competency-based TVET systems? Literature was identified through Google Scholar, Scopus, ERIC, ScienceDirect, and IEEE Xplore. Additional policy and institutional sources were drawn from international organizations and Zambian TEVET documents where they were directly relevant to digital transformation, vocational learning, competency-based training, or institutional implementation. Search terms included combinations of "emerging technologies", "artificial intelligence", "learning analytics", "simulation", "extended reality", "digital assessment", "technology adoption", "TVET", "vocational education", "competency-based training", "occupational competence", "digital pedagogy", and "institutional readiness". Sources were included where they addressed emerging technologies in TVET or vocational education; explained technology adoption, socio-technical implementation, or institutional readiness; examined vocational pedagogy or competency-based learning; discussed AI, simulation, XR, learning analytics, or digital assessment; or provided relevant TVET or TEVET policy guidance. Foundational theoretical works were retained where they offered key explanatory lenses. At the same time, recent sources from 2024 to 2026 were prioritized because they address current developments in AI, XR, analytics, digital assessment, and technology-supported vocational training. Sources were excluded where they focused only on general education without clear relevance to vocational learning, discussed technology without linking it to pedagogy or competence development, or lacked sufficient conceptual or methodological detail. The analysis was conducted through thematic synthesis. Sources were first grouped according to their main explanatory focus: technology adoption, socio-technical implementation, pedagogical learning theory, teacher knowledge, digital transformation, competency-based assessment, and institutional readiness. Each perspective was then examined in terms of what it explains, what it overlooks, and how it contributes to understanding technology-mediated competence development in TVET. This process led to the four analytical dimensions used in the paper: technological affordances, pedagogical design, institutional readiness, and competency-based assessment.

III. THEORETICAL PERSPECTIVES ON EMERGING TECHNOLOGY ADOPTION

Emerging technology adoption in vocational education is often discussed in terms of infrastructure, digital tools, and willingness to use technology. While these issues are important, they do not fully explain adoption in competency-based TVET systems. In TVET, technology adoption is also shaped by institutional policy, educator capability, assessment requirements, industry expectations, technical support, and the organization of practical learning environments. Recent literature on AI and immersive technologies reinforces this point. AI is increasingly considered in relation to curriculum development, qualifications, labor-market responsiveness, learner support, and assessment [18]. Mixed reality, including virtual and augmented reality, has also shown potential to support behavioral, cognitive, and affective training outcomes in vocational education [19]. However, these technologies raise questions about governance, data quality, transparency, equity, assessment validity, instructional design, and transfer to workplace performance [18], [19]. They should therefore be judged not by innovation value alone, but by how well they support the development of credible competence. Technology adoption models provide an important starting point because they explain factors that influence whether individuals accept and use new systems. The Unified Theory of Acceptance and Use of Technology identifies performance expectancy, effort expectancy, social influence, and facilitating conditions as key determinants of technology use [9]. These factors are relevant to TVET because educators and learners are more likely to use emerging technologies when they perceive them as useful, manageable, and institutionally supported. However, adoption models explain user acceptance more clearly than they explain whether a

technology supports valid occupational learning, aligns with competency standards, or fits the realities of workshops, laboratories, and workplaces.

Socio-technical systems theory provides a broader explanation by treating technology adoption as an interaction between technical tools and social systems. Technologies are not neutral additions to existing practice; they become part of institutional routines, roles, infrastructure, policy arrangements, and quality assurance processes [10], [11]. This is important in TVET, where digital tools must operate within systems of curriculum design, assessment, certification, and industry engagement. A simulation tool, learning analytics dashboard, or AI-supported feedback system may work in a pilot, but fail to scale if institutions do not address connectivity, technical support, staff development, procurement, data protection, and assessment procedures [6], [18]. Activity theory adds a further perspective by explaining tensions that emerge when new tools enter established learning systems. It views human activity as occurring within systems composed of subjects, tools, rules, communities, objects, and divisions of labor [20]. In TVET, the objective of the activity is not simply course completion but occupational competence. Emerging technologies may therefore create contradictions between digital learning and existing rules for assessment, workshop practice, timetabling, institutional support, or workplace supervision [21]. These tensions reveal that implementation depends on how the wider activity system adjusts, not merely on whether users accept the technology. Taken together, these perspectives show that the adoption of emerging technologies in TVET is a multi-level process involving user acceptance, institutional alignment, and implementation tensions. However, they remain incomplete without attention to learning theory, because they do not fully explain how learners develop competence through technology-mediated practice.

IV. PEDAGOGICAL PERSPECTIVES ON EMERGING TECHNOLOGY USE

The theoretical perspectives discussed above explain why user acceptance, institutional systems, and implementation dynamics shape the adoption of emerging technologies. However, adoption alone does not guarantee meaningful learning. In vocational education, the central concern is whether technology helps learners develop, practice, and demonstrate occupational competence. Constructivist learning theory views learning as an active process in which learners construct knowledge through engagement with tasks, tools, problems, and feedback [12]. This aligns with competency-based training, where competence develops through repeated practice, reflection, and improvement rather than through the passive reception of information. Emerging technologies can support this process by enabling learners to manipulate digital objects, test decisions, correct errors, and receive timely feedback. Simulation environments, AI-supported practice tasks, and interactive platforms are therefore useful only when designed around meaningful learner activity rather than simple content delivery. Situated learning is also important because vocational competence is closely connected to context. Learners need to apply knowledge in situations that resemble real occupational practice [13]. This perspective is relevant for virtual reality, augmented reality, mixed reality, digital twins, and work-based learning platforms. Recent evidence suggests that immersive technologies can support training outcomes, but their value depends on instructional design and connection to training goals [19]. A simulation that appears realistic but is poorly aligned with occupational standards may engage learners without developing valid competence. Communities of practice, connectivism, and TPACK add social, networked, and teacher-focused dimensions. Vocational learning often takes place through interaction with peers, lecturers, workplace supervisors, and practitioners [14]. Digital portfolios, online collaboration, remote mentoring, and peer feedback can extend these interactions, while connectivism highlights the need for learners to navigate changing networks of people, information, and digital systems [22], [23]. However, access to information is not the same as competence; learners still require structured practice, guided feedback, and performance against standards. TPACK further shows that effective integration requires the interaction of technological, pedagogical, and content knowledge [24]. In TVET, this means lecturers need to understand not only how a tool works but also how it supports a specific vocational task, learning outcome, and assessment requirement.

Recent guidance on AI in TVET reinforces this pedagogical concern by emphasizing educational purpose, human agency, equity, safety, and responsible institutional practice [6]. AI should support the lecturer's judgment and practical training rather than replace them. Learning analytics may identify learner difficulties, and automated feedback may support formative learning, but neither confirms occupational competence without valid assessment criteria and human oversight. Taken together, these perspectives show that emerging technologies should not be judged by novelty or technical sophistication alone. Their usefulness depends on whether they support active learning, authentic practice, social participation, guided feedback, lecturer capability, and valid evidence of competence.

V. SYNTHESIS OF THEORETICAL AND PEDAGOGICAL PERSPECTIVES

The reviewed literature shows that a single theory cannot adequately explain the adoption of emerging technologies in TVET. Adoption models explain why educators and learners may accept or resist technology, but they say less about the quality of vocational learning that follows. Socio-technical perspectives explain the relationship among tools, institutions, and organizational routines, but they do not fully account for how learners develop occupational competence. Pedagogical theories explain learning processes, but they are often less explicit about infrastructure, policy, quality assurance, and institutional readiness. A more useful explanation, therefore, emerges when these perspectives are considered together. The synthesis developed in this paper positions the adoption of emerging technologies in TVET as an alignment issue. In competency-based vocational education, digital tools should not be judged solely by their availability, novelty, or technical sophistication. Their value depends on how well they align with pedagogical design, institutional readiness, and competency-based assessment. This alignment is especially important for technologies such as artificial intelligence, learning analytics, simulation, extended reality, and digital platforms, because these systems may influence not only teaching and learning but also feedback, learner support, evidence generation, and assessment decisions [6], [18], [19].

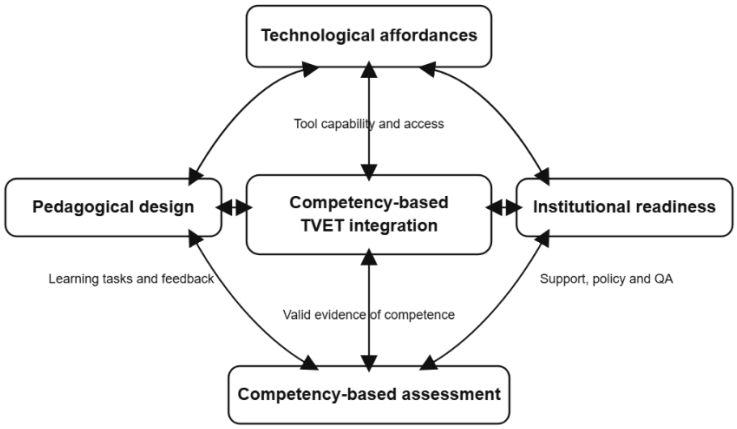


Figure 1. Conceptual alignment framework for emerging technology adoption in TVET

The synthesis is further summarized in Table 1.

Table 1. Dimensions For Analyzing Emerging Technology Adoption In TVET

Dimension	Main focus	Contribution to TVET	Risk if weakly aligned

Technological affordances	Tool maturity, accessibility, functionality, and educational usefulness	Supports practice, feedback, simulation, learner support, and evidence capture	Novelty is prioritized over competence
Pedagogical design	Learning tasks, feedback, authentic participation, and lecturer mediation	Links technology to active, situated, and competence-oriented learning	Engagement does not translate into occupational skill
Institutional readiness	Infrastructure, leadership, staffing, technical support, policy, and quality assurance	Enables implementation, scaling, maintenance, and sustainability	Pilots fail because support systems are weak
Competency-based assessment	Occupational standards, performance evidence, assessment validity, and certification credibility	Links digital learning to a credible demonstration of competence	Platform activity is mistaken for occupational performance

Table 1 shows that each dimension contributes to technology integration, but also introduces risks when weakly aligned. Technological affordances matter because tools must support practical skill development, feedback, and evidence capture rather than novelty alone. Pedagogical design matters because digital tools only strengthen TVET when they support active, situated, and competence-oriented learning. Institutional readiness matters because technologies require infrastructure, leadership, technical support, staff development, and quality assurance to move beyond pilots. Competency-based assessment matters because platform activity, simulation completion, or AI-generated feedback cannot be treated as evidence of competence unless linked to valid occupational standards and performance evidence. The framework, therefore, shifts the discussion from technology adoption to technology integration for competence. The Zambian TEVET context reinforces this wider lesson. Because TEVET combines technical, vocational, and entrepreneurship training, technology adoption should support not only the acquisition of technical skills but also problem-solving, innovation, employability, and credible assessment. Emerging technologies are therefore most useful when they strengthen the relationship between learning tasks, occupational standards, institutional support, and evidence of competence [16], [17].

VI. DISCUSSION AND IMPLICATIONS

The synthesis presented in this paper suggests that the adoption of emerging technologies in TVET should be understood as a process of educational, technical, and institutional alignment. The central issue is not whether a technology is modern or available, but whether it supports credible skills development within a competency-based system. Adoption decisions should therefore begin with occupational tasks, learning outcomes, and assessment requirements before moving to tools and platforms. Different types of digitalization contribute to TVET in different ways. Digital learning platforms support access to content, communication, and flexible delivery; learning analytics can support learner tracking and early intervention; simulation and extended reality can support practical rehearsal and safe experimentation; and artificial intelligence can support feedback, adaptive learning, resource development, and learner support. However, each technology also introduces risks. Platforms may encourage content delivery without practice; analytics may produce data without meaningful pedagogical interpretation; simulations may appear authentic without matching occupational standards; and AI may generate outputs that require human verification and ethical oversight.

For institutions, pedagogy and assessment, emerging technologies should not be treated as isolated classroom innovations or equipment projects. Their value depends on whether they meet curriculum requirements,

lecturers' capabilities, workshop routines, learners' needs, technical support, and quality assurance arrangements [10], [11]. They should extend rather than replace vocational teaching, workshop practice, and workplace learning. Digital activity should also not be mistaken for occupational performance. Assessment in technology-mediated TVET should still focus on valid evidence of competence, including practical observation, digital portfolios, simulation records, workplace evidence, and lecturer judgment. Where AI or analytics are used, human oversight, transparency, and accountability remain necessary [6], [18]. The Zambian TEVET context illustrates these implications because institutions operate within a system shaped by competency-based training, institutional accreditation, quality assurance, and open, distance, and flexible learning. Emerging technologies should therefore support the credibility of practical training and assessment, rather than simply expanding digital access. For Zambia and similar contexts, the priority is purposeful integration of selected technologies that strengthen learning, inclusion, lecturer support, institutional capacity, and evidence of competence. The proposed framework should be understood as a conceptual and literature-derived synthesis rather than an empirically validated model. Future research should test and refine it using evidence from learners, lecturers, institutional leaders, technical support staff, regulators, and industry partners. Further studies should also examine how AI-supported feedback, learning analytics, simulation, and extended reality affect lecturer readiness, learner experience, assessment validity, cost, sustainability, and transfer to workplace performance.

VII. CONCLUSION

This paper examined the adoption of emerging technologies in TVET from theoretical and pedagogical perspectives. It argued that adoption cannot be understood only as the introduction or acceptance of digital tools. In competency-based vocational education, emerging technologies must be judged by how well they support learning design, institutional implementation, and valid demonstration of occupational competence. The main contribution of the paper is the argument that the adoption of emerging technologies in TVET should be understood as an alignment issue involving technological affordances, pedagogical design, institutional readiness, and competency-based assessment. For TVET systems, including Zambia's TEVET context, digital transformation should begin with the question of which competencies learners are expected to develop, how they will be taught, how they will be assessed, and what institutional conditions are needed to support implementation. Future empirical research should test the framework in specific TVET and TEVET contexts to establish how emerging technologies can strengthen the development of credible competence.

ACKNOWLEDGEMENT

The development of this paper was informed by doctoral supervisory guidance and the institutional context of the Technical and Vocational Teachers' College.

REFERENCES

[1]. International Labor Organization, Shaping Skills and Lifelong Learning for the Future of Work, Report VI, International Labor Conference, 109th Session. Geneva, Switzerland: International Labor Office, 2021. https://www.ilo.org/resource/conference-paper/ilc/109/shaping-skills-and-lifelong-learning-future-work

[2]. S. Jeon, *How Can Innovative Technologies Transform Vocational Education and Training? Insights for Ukraine*. Paris, France: OECD Publishing, 2025. https://doi.org/10.1787/fb40f416-en

[3]. UNESCO, *Transforming Technical and Vocational Education and Training for Successful and Just Transitions: UNESCO Strategy 2022–2029*. Paris, France: UNESCO, 2022. https://doi.org/10.54675/EUDU5854

[4]. J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computers & Education*, vol. 147, Art. no. 103778, 2020. https://doi.org/10.1016/j.compedu.2019.103778

[5]. D. F. Angel-Urdinola, C. Castillo-Castro, and A. Hoyos, "Meta-analysis assessing the effects of virtual reality training on student learning and skills development," Policy Research Working Paper 9587. Washington, DC, USA: World Bank, 2021. https://openknowledge.worldbank.org/entities/publication/77a74b0e-7030-5e69-bd90-49a94f244671

[6]. C. Yang, W. Wu, and H. Tegelbeckers, *Integrating AI in TVET: A Practical Guide for Institutions*. Paris, France, and Bonn, Germany: UNESCO and UNESCO-UNEVOC, 2026. https://doi.org/10.64043/BOWG1644

[7]. S. Billett, *Vocational Education: Purposes, Traditions and Prospects*. Dordrecht, The Netherlands: Springer, 2011. https://doi.org/10.1007/978-94-007-1954-5

[8]. M. Mulder, "Conceptions of professional competence," in *International Handbook of Research in Professional and Practice-Based Learning*, S. Billett, C. Harteis, and H. Gruber, Eds. Dordrecht, The Netherlands: Springer, 2014, pp. 107–137. https://doi.org/10.1007/978-94-017-8902-8_5

[9]. V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003. https://doi.org/10.2307/30036540

[10]. F. W. Geels, "Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study," *Research Policy*, vol. 31, no. 8–9, pp. 1257–1274, 2002. https://doi.org/10.1016/S0048-7333(02)00062-8

[11]. W. J. Orlikowski, "Using technology and constituting structures: A practice lens for studying technology in organizations," *Organization Science*, vol. 11, no. 4, pp. 404–428, 2000. https://doi.org/10.1287/orsc.11.4.404.14600

[12]. D. H. Jonassen, "Designing constructivist learning environments," in *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory*, C. M. Reigeluth, Ed. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1999, pp. 215–239. https://doi.org/10.4324/9781410603784-12

[13]. J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*. Cambridge, U.K.: Cambridge University Press, 1991. https://doi.org/10.1017/CBO9780511815355

[14]. E. Wenger, *Communities of Practice: Learning, Meaning, and Identity*. Cambridge, U.K.: Cambridge University Press, 1998. https://doi.org/10.1017/CBO9780511803932

[15]. S. Rajamanickam, R. Che' Rus, and M. N. Abdul Raji, "Enhancing TVET for a digital-ready workforce: A systematic literature review," *International Journal of Modern Education*, vol. 6, no. 23, pp. 865–881, 2024. https://doi.org/10.35631/IJMOE.623060

[16]. Republic of Zambia, *Technical Education, Vocational and Entrepreneurship Training Act*. Lusaka, Zambia: Government Printer, 1998. https://zambialii.org/akn/zm/act/1998/13

[17]. Republic of Zambia, *Technical Education, Vocational and Entrepreneurship Training Policy*. Lusaka, Zambia: Ministry of Higher Education, 2019. https://www.teveta.org.zm/archive/downloads/1607525522.pdf

[18]. OECD, *Developing Vocational Education and Training with Artificial Intelligence*. Paris, France: OECD Publishing, 2026. https://doi.org/10.1787/e9f76b4e-en

[19]. R. Bödding, S. A. Schriek, and G. W. Maier, "A systematic review and meta-analysis of mixed reality in vocational education and training: Examining behavioral, cognitive, and affective training outcomes

and possible moderators," *Virtual Reality*, vol. 29, Art. no. 44, 2025. https://doi.org/10.1007/s10055-025-01118-z

[20]. Y. Engeström, *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Helsinki, Finland: Orienta-Konsultit, 1987. https://lchc.ucsd.edu/mca/Paper/Engestrom/Learning-by-Expanding.pdf

[21]. V. Kaptelinin and B. Nardi, *Acting with Technology: Activity Theory and Interaction Design*. Cambridge, MA, USA: MIT Press, 2006. https://mitpress.mit.edu/9780262513319/acting-with-technology/

[22]. G. Siemens, "Connectivism: A learning theory for the digital age," *International Journal of Instructional Technology and Distance Learning*, vol. 2, no. 1, pp. 3–10, 2005. https://www.urbansystemichealthpartnership.org/s/connectivism.pdf

[23]. S. Downes, *Connectivism and Connective Knowledge: Essays on Meaning and Learning Networks*. Ottawa, Canada: National Research Council Canada, 2012. https://www.oerknowledgecloud.org/archive/Connective_Knowledge-19May2012.pdf

[24]. P. Mishra and M. J. Koehler, "Technological pedagogical content knowledge: A framework for teacher knowledge," *Teachers College Record*, vol. 108, no. 6, pp. 1017–1054, 2006. https://doi.org/10.1111/j.1467-9620.2006.00684.x

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Randomized Backtracking: Another Unusable Sampling Method for the Traveling Tournament Problem?

Timothy Böhm
Department of Computer Science, VU University
Amsterdam, The Netherlands
t.boehm@student.vu.nl
0009−0004−8550−5805

Daan van den Berg
Department of Computer Science, VU University
Amsterdam, The Netherlands
daam@yamasan.nl
0000−0001−5060−3342

Bas Loyen
Department of Computer Science, VU University
Amsterdam, The Netherlands
s.m.g.loyen@student.vu.nl
0009−0002−7935−4065

Abstract
The Traveling Tournament Problem is NP-hard, meaning it cannot be solved in polynomial time. For these problems, metaheuristics often provide usable alternatives. These algorithms, however, need one or more valid initial solutions, and as it stands, there is no algorithm to supply them admissibly. In this paper, we explore whether a randomized backtracking algorithm could provide uniform random initial solutions to facilitate the deployment of metaheuristics to the TTP.

Keywords: Traveling Tournament Problem, Sampling Algorithms, Metaheuristics

I. INTRODUCTION

The Traveling Tournament Problem (TTP) involves organizing a tournament among an even number of teams (n_{teams}) on a fixed schedule of $2 \cdot n_{teams} - 2$ rounds (see Figure 1). In the tournament, each team must face every other team exactly twice — once at home and once away — a pattern known as the **double round-robin constraint**.

A second constraint concerns the teams' itineraries: the **maxStreak** condition limits the number of consecutive home or away games a team may play. The standard setting is maxStreak = 3, meaning a team cannot play more than three consecutive home or away games [1].

In addition, the schedule must satisfy the **noRepeat** constraint. If team A plays team B in a given round, the reverse fixture (B hosting A) is not allowed to occur in the immediately following round. This constraint can be generalized to a parameterized form, denoted as noRepeat = n, meaning that a rematch in reversed home/away order must be separated by at least n rounds [2]. Still, in nearly all existing literature, as in this study, this parameter is set to noRepeat = 1.

The objective of the TTP is to construct a schedule that minimizes the total distance traveled by all teams [3]. The problem is NP-hard [4], implying that unless P = NP, which is very unlikely [5], no polynomial-time exact algorithm is currently known and is unlikely to exist. This in itself is bad for problem-solving objectives, but the TTP has an even more pernicious characteristic: it cannot be sampled. This is bad. Other NP-hard problems, like the partition problem [6–8] or the traveling salesman problem [9], [10], can to some degree be solved by metaheuristics, a class of algorithms that run in “anytime complexity,” meaning the user can start, stop, or pause the algorithmic run at any time and take the current solution as definitive. This is a good alternative to the inadmissibility of exponential-time exact algorithms, but the drawback is that metaheuristic algorithms offer no guarantee of solution quality. But generally speaking, solutions get better as the algorithm runs longer, and outcomes for many problem instances are reasonably good.



Figure 1. **Left:** A valid 4-team TTP tournament. **Right:** An invalid TTP tournament, having a noRepeat violation in column 3, a maxStreak = 3 violation in column 2, and doubleRoundRobin violations in rounds (rows) 2, 3, and 4.

For many practical applications, however, these metaheuristic algorithms are a blessing, as their solution quality, though not perfectly optimal, is still much better than any non-computational performance. But metaheuristic algorithms also have a few modest operational requirements. Before starting the run, one or more valid initial solutions to improve upon must be provided. To this end, the algorithm needs a sampling method that produces valid initial solutions, preferably uniformly sampled from the solution space, to avoid biasing the subsequent run. Second, one needs a mutation operator that mutates one valid solution into another and, ideally, also connects all valid solutions into a single large neighborhood while running in polynomial time [2]. A third, optional requirement is a crossover operator that combines two valid individuals to produce a new valid individual. As it turns out, the traveling tournament problem has none of these. Despite a large number of metaheuristic applications on the TTP (see Section 2 in Hassel et al. [2]), none of the studies present an admissible method for sampling, mutation, or crossover operators. In this paper, we will introduce a new sampling approach to the problem and quantify the uniformity of its outcomes. We tentatively define an **admissible sampling method** to

1. Pick one valid solution from the set of all possible solutions to a problem with uniform probability;
2. run in deterministic polynomial time.

At the current state of technology, an admissible sampling method for the traveling tournament appears not to exist. An alternative, however, can be provided by constructive algorithms such as the polygon method [11] or the improvement thereof reported by Nakahata et al. [12]. Both of these algorithms run in deterministic polynomial time and produce only valid schedules, but the outcomes are exclusively *regular*, thereby representing only a tiny region of the solution space.



An alternative is the “rows-first” method, which uniformly generates random matchups for one round after another. Upon its inception, this method was expected to produce uniform samples; in this paper, we will, for the first time, provide evidence that it indeed does (see Section 3 and Figure 2).

Rows-first has a counterpart method in “columns-first”, which generates a full match list for one team (or “column” in the schedule of Figure 1). The advantage of this method over rows-first is that it eliminates half of the noRepeat violations, as reported in Verduin et al. [13]. Still, these authors appear to forget that although the number of violations drops for known constraints, the columns-first method also introduces a *new* type of constraint violation: mismatched teams within a single round.

Although rows-first and columns-first themselves run in deterministic linear time, the produced schedules do contain violations. Once we require *valid* schedules from the algorithm, forcing it to resample whenever an invalid solution is reached, it runs something that looks like stochastic exponential time [14]. We won't provide evidence for this observation; instead, we offer a budget comparison by matchup as a suggestion for future quantitative studies. We also suggest a way of measuring the uniformity of the outcomes.

As a last novelty, we introduce “randomized backtracking,” a new sampling algorithm for the traveling tournament problem. It is a depth-first recursive algorithm based on row-first, but backtracks upon constraint violations. Even though backtracking sampling could be easily considered inadmissible, as it runs in deterministic exponential time (or even deterministic factorial time for some problems), it might not be in practice. This observation, and its sample distribution, make it a surprising addition to the almost empty library of sampling methods for the traveling tournament problem.

The rest of this paper is organized as follows: in the Methods section, we introduce the randomized backtracker and the rows-first algorithm, as well as the methods used to measure performance and assess the uniformity of the resulting distribution. The results are presented in the third section, after which we will discuss the not-so-trivial outcomes in the closing section.

II. METHOD

In this section, we describe two different methods for sampling schedules for nteams = 4, propose a method for measuring uniformity, and provide a way to compare runtimes. First, we sample valid half-normalized TTP schedules with the randomized backtracking algorithm described in [15]. “Half-normalized” in this sense means the first row is always filled out with matchups (1,2) and (3,4) as a method for breaking rotational and permutational symmetry.

The randomized backtracking algorithm recursively places randomized matchups in rows 2, 2, 3, 3, 4, ..., 2(n − 1), checking at each step whether any constraints are violated or will be violated in the future. For example, if all remaining games for a given team are home games, a maxStreak violation is unavoidable. Whenever this happens, the algorithm backtracks immediately, removes the most recently placed matchup from the schedule, and instead tries the next matchup in the randomized list. This process continues until a valid schedule is found (see pseudocode in Algorithm 1).

Algorithm 1 The procedure **DFS-Sampler** recursively tries adding matchups to the current permutation of the schedule. Whenever a constraint is violated it backtracks and tries different matchup, until a valid schedules is found.

```
procedure DFS-SAMPLER(schedule, matchups)
    if schedule is complete then
        return
    end if

    for all m in matchups do
        if CHECKCONSTRAINTS(schedule, m, matchups) then
            continue
        end if

        schedule.append(m)
        matchups.remove(m)                          ▷ Updating current state

        DFS-SAMPLER(schedule, matchups)             ▷ Recursive call

        schedule.remove(m)
        matchups.insert(m)                          ▷ Revert to previous state
    end for
end procedure
```

We are sure the algorithm halts, as it is randomized only in its traversal order and thus remains exhaustive over the huge but finite search space. We are also sure that a valid solution exists, because the TTP with n_teams = 4 has recently been shown to have 160 valid half-normalized schedules [15], [16]. As a final technical note, a new random matchup list is created for each schedule.

Second, we sample valid TTP solutions with a slightly adapted version of rows-first. Like Verduin's classic version [13], the algorithm generates randomized matchups for each row, but immediately terminates whenever a placed matchup violates a constraint and starts anew. This effectively means that a valid schedule will be found only if the randomized matchup lists for each consecutive round do not violate any constraints. This algorithm terminates only stochastically, but in theory, it could run forever without encountering a valid solution. This never happened; however, we sampled 100,000 valid half-normalized schedules for both methods and counted the multiplicity of each unique valid schedule, yielding a frequency histogram (Figure resultsuniform). To determine whether these samples follow a uniformly random distribution, we used Pearson's chi-squared test [17] for an assumed uniform distribution of #samples / #unique schedules = 10^5 / 160 = 625 samples per unique schedule, and calculated p-values. For Pearson’s chi-squared test, we used the default significance level of 5%, as there was no compelling reason to deviate from this.

III. RESULTS

It turns out that the randomized backtracking method does not uniformly randomly sample valid schedules at all (Figure 2). Pearson's chi-squared test confirms this with p ≈ = 0.0 for the uniform distribution. Moreover, the most and least frequently sampled schedules occur 59 and 2,027 times, respectively, illustrating the tremendous non-uniformity of the proposed algorithm. To find these 100,000 schedules, the randomized backtracker required approximately 70.5 placements per schedule, or 7,045,027 placements in total, which, in our implementation, took only a few seconds on 128 cores.

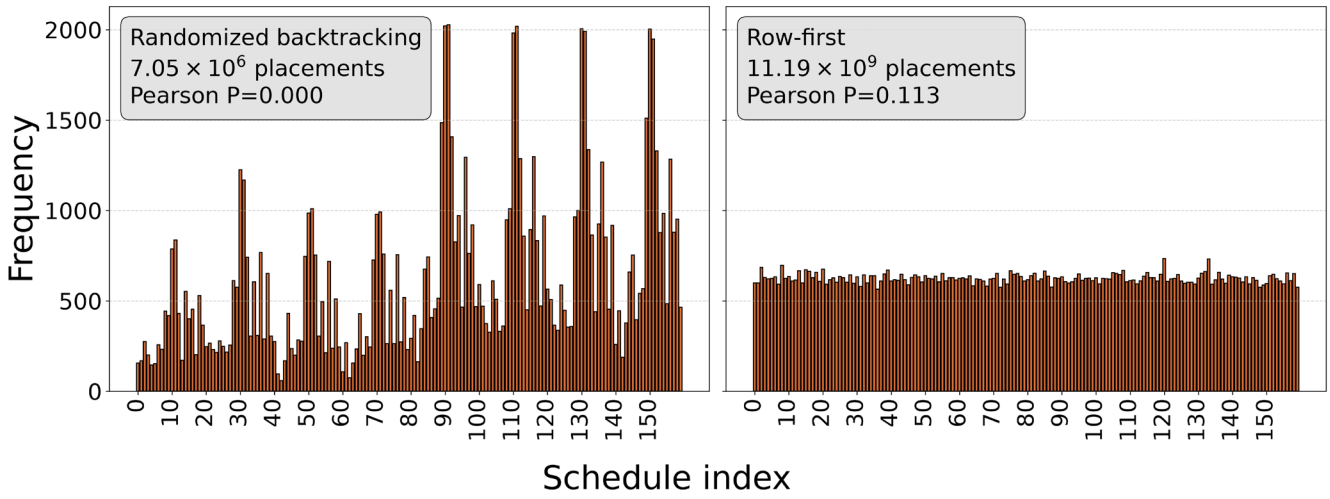


Figure 2. Frequency of results for the generation of 100,000 valid TTP-schedules of size $nteams = 4$ by two algorithms. On the right, the rows-first algorithm yields an almost uniform distribution, whereas the randomized backtracking yields an erratic distribution. Schedule indices on the horizontal axes are the same in both subfigures.

We also ran the randomized backtracker on 8 nodes, each with 128 cores, for nteams = 6, producing 100 million schedules in approximately 2 minutes. For $n_{teams} = 8$, 100 million schedules were found in approximately 3 hours on the same system. These results, however, could not be used in this study for two reasons: first, the problem has 54,427,728 valid schedules for $n_{teams} = 6$, and presumably several billion for $n_{teams} = 8$. These numbers are too large to save to disk comfortably, let alone use to construct a useful histogram. We do, however, report that while exploring the options for this study, we found 702,960,132 unique valid schedules for $n_{teams} = 8$, which is a new lower bound. Second, rows-first did not find a single valid schedule for $n_{teams} = 6$ or $n_{teams} = 8$ in 8 hours on 8 nodes, each with 128 cores, making a distribution comparison impossible. Finding not even a single valid solution for such small instance sizes paints a very bleak future for this approach to TTP sampling, even though it is uniform.

For rows-first sampling, the distribution outcome was quite different. All samples occur roughly #samples / #unique schedules = 10^5 / 160 = 625 times, which is the number we would expect under a uniform distribution. Pearson's chi-squared test confirms that the sample frequencies follow a uniform distribution, with p ≈ = 0.113. The most frequently sampled schedule was seen 735 times, while the least frequently sampled schedule occurred 565 times. When constraint violations are ignored, rows-first runs in linear time. Still, when restarting upon invalid schedules, rows-first requires at least 11,194,726,501 matchup placements distributed across 4,464,072,865 restarts, averaging approximately 2.51 placements per restart, to generate these 100,000 samples. This process took approximately 3 minutes, but with an average of 11,195 matchup placements to



obtain one valid schedule, it is approximately 156 times more expensive than randomized backtracking — and that fraction might grow for higher numbers of n teams.

IV. CONCLUSION & DISCUSSION

The observations in this study turn the TTP's sampling problem on its head. First of all, it seems we have finally found a way to uniformly and randomly sample TTP solutions. Rows-first does the job, but, weirdly enough, despite its "linearity," it takes much more time than the randomized backtracker, which is naturally exponential in complexity. The big question is how these algorithms scale, but because rows-first returned zero results for n_teams = 6 and n_teams = 8, this does not exactly look appealing. However, there is still room for improvement in this method, such as checking before placing or evaluating certain constraint violations (e.g., DRR) only after completion of the schedule. For random backtracking, there might also be ways forward. Even though it does not appear to produce uniform solutions, it does yield solutions with some degree of randomness, which could be seen as an improvement over previous work by DeWerra and Nakahata [11], [12]. But their methods run in linear time, and random backtracking theoretically does not, so the real question is how it scales in time. Regarding the non-uniformity of the results (Figure 2), it is worth noting that the order of schedules might not be preserved across trials. In other words, if we generate a million schedules twice, the most frequent schedule might not be the same in both trials, even though the distributions are nearly identical. At this point, however, this is a mere suspicion, and rigorous testing into these and other details of the distribution should clarify the matter. Two other aspects warrant further inspection. First of all, it might be possible to correct the non-uniformity in some formulaic way. Second, the row order for matchup placement, 2, 2, 3, 3, ..., 5, 6, 6 in this study, might also be randomized. The effects on runtime, measured in matchup placements or wall-clock time, and on uniformity, as calculated using Pearson's correlation coefficient, are of interest. But the weirdest suggestion this study makes is that sampling algorithms might "switch sides", from polynomial to exponential or the other way around, when used for random sampling. Furthermore, it could be true that uniformity comes at the cost of once-linear-now-exponential time complexity, while relinquishing that property might reward the programmer with an exponential-time algorithm behaving polynomially. Finally, reviewers of IHCONCS suggested hybrid reinforcement learning [18], deep generative neural networks, and diffusion models to avoid constraint violations. This is a deluge of ideas whose merit should be investigated in future work.

V. CREDIT STATEMENT AND DISCLAIMER

Timothy Böhm: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Review & Editing, Visualization. **Bas Loyen**: Methodology, Software, Validation, Writing - Original Draft, Writing - Review & Editing. **Daan van den Berg**: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration. Python and C++ source code and auxiliary materials for this publication are publicly available [19]. The logos in this figure were (re)made by melling2293@Flickr and are distributed under Creative Commons license.

REFERENCES

[1]. Nikolaus Frohner, Bernhard Neumann, and Günther Raidl. A beam search approach to the traveling tournament problem. In Evolutionary Computation in Combinatorial Optimization, pages 67–82. Springer International Publishing, 2020.

[2]. Lukas Hassel, Guilherme Nakahata, Ruben Horn, and Daan van den Berg. The traveling tournament problem: Constraint violations for different values of noRepeat and maxStreak. In Proceedings of the

18th International Joint Conference on Computational Intelligence ECTA, 2026. To appear in October 2026.

[3]. Kelly Easton, George Nemhauser, and Michael Trick. The traveling tournament problem description and benchmarks. In Principles and Practice of Constraint Programming—CP 2001: 7th International Conference, CP 2001, Paphos, Cyprus, November 26–December 1, 2001, Proceedings 7, pages 580–584. Springer, 2001.

[4]. Clemens Thielen and Stephan Westphal. Complexity of the traveling tournament problem. Theoretical Computer Science, 412(4–5):345–351, 2011.

[5]. William I. Gasarch. Guest column: The second P = NP poll. ACM SIGACT News, 43(2):53–77, 2012.

[6]. Nikita Sazhinov, Ruben Horn, Pieter Adriaans, and Daan van den Berg. The partition problem and how the distribution of input bits affects the solving process. In Proceedings of the 15th International Joint Conference on Computational Intelligence ECTA, pages 143–150, 2023.

[7]. Ruben Horn, Reitze Jansen, Okke van Eck, and Daan van den Berg. Separating the yes from the no instances in the number partitioning problem. In 16th International Joint Conference on Computational Intelligence, IJCCI 2024, pages 181–188. SciTePress, 2024.

[8]. Daan van den Berg and Pieter Adriaans. Subset sum and the distribution of information. In Proceedings of the 13th International Joint Conference on Computational Intelligence ECTA, pages 134–140, 2021.

[9]. Xin-Ai Dou, Qiang Yang, Xu-Dong Gao, Zhen-Yu Lu, and Jun Zhang. A comparative study on crossover operators of the genetic algorithm for the traveling salesperson problem. In 2023, the 15th International Conference on Advanced Computational Intelligence (ICACI), pages 1–8. IEEE, 2023.

[10]. Darrell Whitley, Doug Hains, and Adele Howe. A hybrid genetic algorithm for the traveling salesperson problem using generalized partition crossover. In International Conference on Parallel Problem Solving from Nature, pages 566–575. Springer, 2010.

[11]. D. de Werra. Some graph models for scheduling sports competitions. Discrete Applied Mathematics, 21(1):47–65, 1988.

[12]. Guilherme Nakahata, Florian Richoux, Daan van den Berg, and Claus Aranha. A constructive method for generating many valid initial solutions to the traveling tournament problem. In International Conference on the Applications of Evolutionary Computation (Part of EvoStar), pages 441–456. Springer, 2026.

[13]. Kristian Verduin, Ruben Horn, Okke van Eck, Reitze Jansen, Thomas Weise, and Daan van den Berg. The traveling tournament problem: Rows-first versus columns-first. In Proceedings of the 26th International Conference on Enterprise Information Systems, ICEIS 2024, volume 1, pages 447–455, 2024.

[14]. Kristian Verduin, Sarah L. Thomson, and Daan van den Berg. Too constrained for genetic algorithms. Too hard for evolutionary computing. The traveling tournament problem. In Proceedings of the 15th International Joint Conference on Computational Intelligence ECTA — Volume 1, pages 246–257, 2023.

[15]. Bas Loyen, Duncan Bart, Florian Richoux, and Daan van den Berg. The traveling tournament problem: Valid solutions are different across instance sizes. In International Joint Conference on Computational Intelligence, pages 491–503. Springer, 2025.

[16]. Bas Loyen, Lukas Hassel, Sarah L. Thomson, Thomas Weise, and Daan van den Berg. The traveling tournament problem: Valid solutions vary widely in LBAs of EvoSTAR 2025, 2025.

[17]. Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302):157–175, 1900.

[18]. Zied Ben Hazem. A fuzzy-TD3 hybrid reinforcement learning framework for robust trajectory tracking of the Mitsubishi RV-2AJ robotic arm. Scientific Reports, 16, Article 12269, 2026.

[19]. Anonymous. Repository containing source material: https://github.com/OdyMeister/TTP-Uniformity, 2025.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Educational Research in Computer Science: Data Collection, Processing, and Analysis

Carolin Hannusch
Department of Computer Science Faculty of Informatics University of Debrecen Debrecen, Hungary
hannusch.carolin@inf.unideb.hu
0000-0002-0098-7293

Mária Csernoch
Department of Computer Science Faculty of Informatics University of Debrecen Debrecen, Hungary
csernoch.maria@inf.unideb.hu
0000-0002-7088-7714

Abstract
This work explores the complete process of educational research in the computer sciences, beginning with classroom-based inquiry in elementary schools and extending through data collection, statistical processing, and analysis. It provides a practical overview of how raw observations are transformed into meaningful insights, highlighting methodologies, tools, and best practices for conducting and interpreting research in educational settings. In this chapter, we follow a research case study from start to finish. From preparation (including software development) and data collection to data analysis and conclusion derivation, we experience the entire research process. What is unique about this chapter is that we address the "unwritten" aspects of field research, including difficulties and challenging situations, and how we addressed them.

Keywords: seducation, text management, digital waste, field research, statistical analysis

I. INTRODUCTION

Digital text management (DTM) is one of the most common activities in the digital world. Yet it is often misinterpreted as synonymous with word processing, learning, teaching [1], and the use of word processors [2]. This misconception is reinforced by assumptions that everyone can perform these tasks without prior knowledge [3], that software tools alone provide sufficient problem-solving skills [4], [5], and that digital immigrants cannot teach digital-born students [6], [7], [8]. Schools often treat DTM as one of the simplest digital activities (e.g, [9]), overlooking both the cognitive load it entails [10], [11], [12] and the need to assess the effectiveness of students and teachers [13], [14]. In practice, digital texts (DTs) collected from students, teachers, and the internet reveal widespread inefficiencies, leading to substantial, often unnoticed waste in the form of ineffective use of time and resources [15], [16]. As Wolfram ([17] p.272) observes, "with no effective, general education in computational thinking, most people can easily be misled, and they are." Today's communication often resembles "computational grunts, not a universal computational language" [8]. The complexity of DTs, the limitations of working memory, and systemic distortions such as cascade noise, group noise, and group polarisation can allow errors to go unnoticed and uncorrected, thereby multiplying waste [18]. Contributing factors include folk-pedagogy [19], teachers' belief in the fixed nature of science [20], the denial of ignorance [18], and the Dunning–Kruger effect [21]. Addressing these problems [16] requires establishing standards in DTM. In other domains, such as production, services, construction, or administration, the lack of standards is recognized as a major source of waste and a barrier to development [22], [23]. Our research group has compiled resources to support standardization and digital sustainability in DTM, thereby improving learning and teaching processes, end-user activities, and the quality of digital artifacts [24]. To put these resources into practice, we designed a field research experiment to examine their applicability in real educational and professional contexts. The experiment was conducted with multiple groups of participants to gain insight into diverse perspectives on DTM standardization. By observing authentic work processes, we wanted to evaluate both user-friendliness and performance impact. This approach allows us to link theoretical concepts with the practical experiences of

professionals. In the next section, we present the methodology and key findings of the field research. We describe our field research, particularly the experiment we conducted across several communities. We focus on our field research methods, developed software, and methodological analysis, with a focus on challenging situations.

II. THE EXPERIMENT

1. Preparation

Our experiment began with the idea of a text management task and the measurement of people's ability to solve it. But how should we measure these abilities? First, we needed a method to obtain analyzable data from the measurements. For this purpose, the logging software Atomic Natural Language Input Tracker Application (ANLITA) was developed [25], which stored data collected on the computer while participants solved text management tasks. A log file recorded all mouse clicks and keystrokes, while a video file of the entire activity was created. In addition, the participants saved all the work they had done on document files. Developing the software was time-consuming, but it was worth the effort. With the ANLITA tool, we can make the field research data calculable, comparable, and analyzable.

2. Collecting data

Our field research consisted of two tasks across four documents. Each task was to be done twice: Once in an erroneous document and once in a properly edited document. The experiment was conducted in Hungarian, but it is language-independent, i.e., it can be adapted to any language. The participants came from several institutions, including teachers, administrative staff, professors, and students (both teenagers and older adults) [26]. The four documents given contained texts arranged in different ways. In Task 1, the provided document contains various errors, the most significant of which are layout errors [24]. In contrast to the erroneous document, the properly edited document contained no unnecessary spaces or carriage return characters. The first task was to add a specified expression – 13 characters – into the indicated paragraph. It was ensured that this modification would not disrupt the text's arrangement in the erroneous document. The participants had to rearrange the text using a method of their choice to make it appear well-arranged and free of layout errors. This task had to be performed in both erroneous and well-structured documents. The second task was to change the font size of the specified text. Here, too, care was taken to ensure that changing the font size in the poorly structured document would not destroy its structure. Participants again had to achieve a properly edited text as the solution to the task. This task was also repeated in the well-structured document. If a participant completed all four tasks, four document files were saved. In addition, the ANLITA software generated a log file containing all the participant's mouse and keyboard activity during the tasks, as well as a video file of the participant's on-screen activity.

III. HANDLING DATA

To capture the dynamics of human-computer interaction, we used ANLITA, our self-developed software platform. ANLITA records keyboard presses and mouse clicks with millisecond accuracy, while simultaneously generating screen video recordings. By combining discrete event logs with continuous visual context, the system creates a rich record of both user input and its immediate effect on the screen. Each event is automatically assigned a unique timestamp, which serves as the basis for synchronizing the different data streams. Because distinct subsystems generate the input logs and video recordings, synchronization is essential. ANLITA embeds reference timestamps directly into the video metadata, enabling input events to be precisely aligned with screen frames. This means that a keystroke or mouse click can be directly mapped to the corresponding visual change

on the display. The synchronized dataset thus enables a detailed reconstruction of the temporal relationship between actions and outcomes. Once collected and aligned, the data is imported into a relational database designed to handle both event-level precision and higher-level queries. The structure is organized around three central components: Events Table: time-stamped keyboard and mouse events, including key codes and cursor positions, Video Metadata Table: references to screen recordings, with frame indices and timestamps for alignment, and a Session Table: contextual information such as session identifiers, user details (if available), and experimental conditions. This schema ensures that individual actions can be linked seamlessly to broader interaction patterns, while also supporting efficient statistical analysis across multiple sessions. Before the data is ready for analysis, several preprocessing steps are applied. First, noise filtering removes incomplete or spurious input events. Next, normalization converts event codes and cursor coordinates into standardized formats. Data is then segmented into meaningful units, such as task phases or fixed time windows, depending on the analytical focus. Finally, aggregated features are computed, including typing speed, click frequency, and idle durations, providing higher-level descriptors that complement the raw event streams. Together, these steps transform raw logs and video data into a structured dataset that is both comprehensive and statistically tractable, forming the foundation for deeper analyses of user behavior and interaction dynamics.

1. Missing data

Dealing with missing data is one of the most difficult challenges in statistical analysis. We wanted to ensure not only a quantitative but also a qualitative analysis of the data collected. For this reason, the ANLITA software package was developed to generate six different output types for all participants:

Task 1: output word file, erroneous document, typing task
Task 2 output word file, erroneous document, font size task
Task 3: output word file, correct document, typing task
Task 4 output word file, correct document, font size task

Logfile containing all keyboard and mouse activities for all tasks

Video file containing everything seen on the monitor while the participant worked on the tasks

In the best-case scenario, each participant generated all six output files. Since our goal was to understand how participants worked, recording their activities on video was crucial for revealing their problem-solving strategies. The logfile contains data on each participant's keyboard and mouse activity. All logfiles combined built the basis for our statistical analysis. Figure 6 shows an extract from such a logfile.

If a participant ended the video recording prematurely or did not start it at all, we did not use that participant’s results. Unfortunately, this method led to some data loss. On the other hand, this ensured that we had a qualitative assignment for each piece of data evaluated. The video recordings had to be viewed manually, one by one, which was very time-consuming but ensured the quality of our results.

IV. ANALYZING DATA

1. SOLO categories

To analyze and correctly classify our results, we use the Structure of the Observed Learning Outcomes (SOLO) categories introduced by Biggs and Collis [27]. The following SOLO categories are used: Prestructural,

Unistructural, Multistructural, Relational, and Extended Abstract. We modified the definitions of each category to ensure they were perfectly tailored to our experiment. First, we needed to categorize the actions that occurred (or should have occurred) while solving the tasks. Table 1 gives the notation of each step.

Table 1..Description of actions in the erroneous/correct document

Action	Description
A	Correcting errors before typing/changing font size
B	Typing / Setting font size
C	Handling multiple Space characters
D	Handling end-of-line Enter characters.
E	Setting left indentation
F	Unnecessary steps (e.g., type – delete – type again)

A different definition of the categories was needed for the solutions on the correct documents (Table 2) and on the erroneous documents (Table 3). For the definitions, we use the usual logical operators AND, OR, and NOT to combine the actions described in Table 1. When working on the correct documents in Task 3, the highest category achieved was 'Unistructural', whereas in Task 4 it was 'Multistructural'.

Table 2. SOLO categories for the correct documents

Category	Description
Prestructural	OR(B,E)
Unistructural	AND(B, NOT(E))
Multistructural	AND(B, NOT(E), F)

In the erroneous documents, more work should have been done. In the optimal case, a participant corrected the errors in the document before they started modifying the text. Therefore, step A is crucial for the highest category 'Extended abstract', as can be seen in Table 3.

Table 3. SOLO categories for the erroneous documents

Category	Description
Prestructural	OR(B, C, D, E)
Unistructural	OR(AND(B, C), AND(B, D), AND(B, E), AND(C, D, E))
Multistructural	OR(AND(B, C, D), AND(B, C, E), AND(B, D, E), AND(C, D, E)
Relational	AND(B, C, D, E, NOT(F))
Extended abstract	AND(A, B, C, D, E, NOT(F))

The SOLO categories help classify the quality of solutions, since, e.g., pushing 10 keyboard presses can lead to a good solution or may not solve a task. To classify each participant's SOLO category, it was crucial to watch

and analyze the videos on the screen. This was time-consuming, but it ensured that the quantitative data were combined with the qualitative characteristics of the solutions.

2. Statistical analysis

The analysis was conducted by carefully examining six output files, each providing unique insights into the participants' work. Wherever possible, automated tools were used to verify the accuracy of the Word documents, identify SOLO categories, and calculate and categorize keystroke and mouse-click counts. However, the numbers alone did not paint a complete picture. To complete the picture, we also used video recordings that revealed finer aspects of performance: whether the "Show/Hide" button was activated at the beginning, how the selection was made - by dragging with the mouse or by more efficient double or triple clicking - and in what order errors were identified and corrected. These details provided valuable insights into the level of understanding and the problem-solving strategies in practice [26]. When examining the patterns contained in the data, we started with simple but effective techniques: counting, calculating averages, and observing deviations. From there, we moved on to more complex areas to compare the work of participants divided into groups. This allowed us to examine not only how much time they spent on the tasks and how they used the keyboard and mouse, but also how they handled errors and which problem-solving strategies they used. Several tools were needed to tie these threads together: Microsoft Excel, SPSS, and R for statistical analysis; video editing software for observing subtle behaviors; and custom applications developed specifically for this study. Together, these methods provided a comprehensive and nuanced picture of how participants solved the tasks, striking a balance between quantitative accuracy and qualitative insight.

V. RESULTS

In this section, we summarise our key findings regarding participants' text management skills and problem-solving strategies. We focus on facts that directly and significantly impact business solutions and affect numerous companies worldwide, as word processing and text-editing skills are important in all types of business communication. In this chapter, we focus on the data we collected in two schools, where participants were divided into three (similar) groups: Grade 7, Grade 8, and Teachers. The results of our statistical analyses are discussed in detail in the research articles [26], [28], and [29]. For our purposes, we will focus on the most meaningful results, both qualitative and quantitative.

1. Qualitative results

Table 4 shows the changes in the participants' SOLO categories between Task 1 and Task 3. As described in Section II, this meant that the same task had to be solved once in a poorly formatted document (Task 1) and once in a well-formatted document (Task 3). The results show that Task 3 imposed a lower cognitive load, leading many participants to produce solutions of much higher quality.

Table 4. Changes in SOLO Categories Task 1 (erroneous) vs Task 3 (correct) typing task

	School 1		**School 2**	
Category	Task 1	Task 3	Task 1	Task 3
P	42.6%	7.9%	48.5%	9.7%
U	44.6%	92.1%	45.2%	90.3%
M	12.9%	—	6,5%	—
R	0.0%	—	0.0%	—
E	0.0%	—	0.0%	—

When examining the changes in SOLO categories between Task 2 and Task 4 – the same font-size task was solved in an erroneous document and then in a correct document (see Section II) – we find a similar pattern, as shown in Table 5. The correct document required less cognitive effort from the participants, as indicated by the relatively low SOLO category levels observed and the higher number of participants achieving better results.

Table 5. Changes in SOLO Categories Task 2 vs Task 4 fontsize setting task

	School 1		**School 2**	
Category	Task 2	Task 4	Task 2	Task 4
P	29.7%	24.8%	26.9%	18.3%
U	42.6%	75.2%	63.4%	81.7%
M	25.7%	0.0%	9.7%	0.0%
R	2.0%	—	0.0%	—
E	0.0%	—	0.0%	—

Another spectacular result is the time saved when working on a well-organized document rather than an erroneous one. Tables 6 and 7 show the differences in average time across groups.

Table 6. Comparison of the average time for Task 1 (err.) vs Task 3 (correct) typing task

	School 1		**School 2**	
	Task 1	**Task 3**	**Task 1**	**Task 3**
Grade 7	154.86 sec	37.14 sec	247.48 sec	53.39 sec
Grade 8	160.16 sec	46.08 sec	187.84 sec	46.59 sec
Teachers	219.82 sec	59.18 sec	167.96 sec	43.96 sec

Table 7. Comparison of the average time Task 2 (err.) vs Task 4 (correct) fontsize setting task

	School 1		School 2	
	Task 2	**Task 4**	**Task 2**	**Task 4**
Grade 7	148.28 sec	52.00 sec	183.94 sec	48.82 sec
Grade 8	138.78 sec	48.95 sec	143.32 sec	62.65 sec
Teachers	160.86 sec	63.32 sec	152.89 sec	51.30 sec

These tables show that all groups in both schools spent much less time on the correct documents (Tasks 3 and 4) than on the erroneous ones (Tasks 1 and 2). Furthermore, the reduction consistently ranged from 60% to 78%.

Considering both the quantitative and qualitative results, we must note that, beyond the documents that required more time, the participants did not complete the tasks in them. None of the participants reached the SOLO categories R or E, which are considered completed. The difference between these two SOLO categories is the effectiveness of the selected problem-solving approach. On the other hand, in the correct documents, the cognitive load imposed on the participants is much lower. This is expressed in the highest SOLO categories. The Unistructural SOLO category indicates that only one aspect of the problem must be recognized, and that is enough to solve the problem. In the font-size task, the highest SOLO category is Multistructural, indicating a more effective solution than Unistructural. It is also remarkable that teachers were not performing better than their students. This implies that teachers lack the level of computational thinking required to develop students' computational thinking skills effectively. This finding indicates that, to develop students' skills, teachers should first be taught to develop their own. However, addressing this problem is beyond the scope of the present paper, which aims only to call attention to it.

VI. CONCLUSION

The research outlined has shown that digital text management, often trivialized as a mere matter of formatting or typing proficiency, is, in fact, a complex cognitive and methodological challenge with far-reaching implications. By following the process from software development through data collection to statistical and qualitative analysis, we not only uncovered how individuals interact with digital texts but also revealed the inefficiencies and systemic barriers that hinder effective work. Our field experiment demonstrated that properly edited documents drastically improve both efficiency and output quality. The correct documents impose less cognitive load on participants, as evidenced by lower SOLO categories; they also require significantly less time to complete the tasks. These results reinforce the necessity of establishing clear standards in digital text management, just as other domains, such as production and services, have long recognized the productivity gains from standardization. These standards were published a long time ago; however, they have not reached a wide audience, which should be one of the aims of digital education. Imagine if all the people around the world who deal with text management as a key part of their work every day, from industrial management to corporate marketing to government ministries, could save about two-thirds of their time. In other words, productivity could be optimized so that, for example, a task that normally takes six hours could be completed in two hours. This presents huge opportunities for companies worldwide. Beyond the quantitative evidence, the qualitative insights highlight important lessons: many participants default to inefficient practices due to entrenched misconceptions, limited training, or systemic neglect of digital literacy as a critical skill. Yet, the improvements we observed across diverse groups demonstrate that better outcomes are not only possible but also highly scalable when supported by appropriate frameworks, methodological clarity, and thoughtful education. In conclusion, the study makes a strong case for rethinking how digital text management is taught, practiced, and valued across

educational and professional contexts. By embracing standardization and sustainability in digital practices, we can reduce waste, enhance problem-solving, and unlock significant gains in productivity and communication quality. The path forward lies not in expecting individuals to adapt to inefficient systems, but in designing environments where effective practices are embedded, supported, and rewarded. In doing so, digital text management can transform from a hidden source of frustration into a foundation for digital excellence.

REFERENCES

[1] V. Woollaston, "The surprisingly subtle ways Microsoft Word has changed how we use language," BBC News, June 20, 2026. [Online]. Available: https://www.bbc.com/future/article/20231025-the-surprisingly-subtle-ways-microsoft-word-has-changed-the-way-we-use-language

[2] M. Csernoch, T. Nagy, and J. Csernoch, "Computer cooking vs problem solving," Teaching Mathematics and Computer Science, vol. 22, no. 1, pp. 35–58, 2024.

[3] M. Csernoch, P. Biró, and C. Hannusch, "Modification of erroneous and correct digital texts," Entropy, vol. 26, no. 12, Art. no. 1015, 2024.

[4] M. Ben-Ari, "Bricolage forever!," in Proc. Psychology of Programming Interest Group, Leeds, U.K., 1999.

[5] M. Ben-Ari and T. Yeshno, "Conceptual models of software artifacts," Interacting with Computers, vol. 18, no. 6, pp. 1336–1350, 2006.

[6] M. Prensky, "Digital natives, digital immigrants," On the Horizon, vol. 9, no. 5, pp. 1–6, 2001.

[7] M. Prensky, "Digital natives, digital immigrants part 2: Do they really think differently?" On the Horizon, vol. 9, no. 6, pp. 1–6, 2001.

[8] P. A. Kirschner and P. De Bruyckere, "The myths of the digital native and the multitasker," Teaching and Teacher Education, vol. 67, pp. 135–142, 2017.

[9] W. Greenwood, "Why Microsoft Word skills are important in the workplace," Priority Management. [Online]. Available: https://www.prioritymanagement.com.au/word-skills-importance/. Accessed: Jun. 1, 2026.

[10] J. Sweller, "Visualization and instructional design," in Proc. Int. Workshop Dynamic Visualizations and Learning, Tübingen, Germany, 2002, pp. 1501–1510.

[11] J. Sweller, P. Ayres, and S. Kalyuga, "Cognitive load theory in perspective," in Cognitive Load Theory, New York, NY, USA: Springer, 2011, pp. 237–242.

[12] O. Kruse et al., Digital Writing Technologies in Higher Education: Theory, Research, and Practice. Cham, Switzerland: Springer Nature, 2023.

[13] J. M. Wing, "Computational thinking," Communications of the ACM, vol. 49, no. 3, pp. 33–35, 2006.

[14] J. M. Wing, "Computational thinking's influence on research and education for all," Italian Journal of Educational Technology, vol. 25, no. 2, 2017.

[15] M. Csernoch, "Teaching word processing: The theory behind," Teaching Mathematics and Computer Science, vol. 7, no. 1, pp. 119–137, 2009.

[16] M. Csernoch, T. Nagy, K. Nagy, J. Csernoch, and C. Hannusch, "Human-centered digital sustainability: Handling enumerated lists in digital texts," IEEE Access, vol. 12, pp. 30544–30561, 2024.

[17] C. Wolfram, The Math(s) Fix: An Education Blueprint for the AI Age. Champaign, IL, USA: Wolfram Media, 2020.

[18] D. Kahneman, O. Sibony, C. R. Sunstein: Noise: A flaw in human judgment. London, U.K.: William Collins, 2022.

[19] R. Lister, "After the gold rush: Toward sustainable scholarship in computing," in Proc. Tenth Australasian Computing Education Conf., Darlinghurst, Australia, 2008, pp. 3–17.

[20] J. Chen, D. Morris, and N. Mansour, "Science teachers' beliefs: Perceptions of efficacy and the nature of scientific knowledge and knowing," in International Handbook of Research on Teachers' Beliefs, New York, NY, USA: Routledge, 2014, pp. 382–398.
[21] J. Kruger and D. Dunning, "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments," Journal of Personality and Social Psychology, vol. 77, no. 6, pp. 1121–1134, 1999.
[22] J. K. Liker, The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer. New York, NY, USA: McGraw-Hill, 2004.
[23] M. Rother, Toyota Kata: Managing People for Continuous Improvement, Adaptiveness, and Superior Results. New York, NY, USA: McGraw-Hill, 2010.
[24] M. Csernoch, "Lean digital education to resolve the paradox of the illusion of digital prosperity," Journal of Innovation and Knowledge, vol. 10, no. 2, Art. no. 100676, 2025.
[25] M. Csernoch and K. Nagy, "Pre-testing erroneous text-based documents: Logging end-user activities," Frontiers in Education, vol. 7, Art. no. 958635, 2023.
[26] M. Csernoch, D. Máté, and S. Kovács, "Ranking cognitive processes for eliminating waste in digital text management," Humanities and Social Sciences Communications, 2026.
[27] J. B. Biggs and K. F. Collis, Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome). New York, NY, USA: Academic Press, 1982.
[28] M. Csernoch and C. Hannusch, "Problem-solving strategies of teachers and students in digital text management," in press.
[29] M. Csernoch, S. Szeghalmy, and P. Biró, "Understanding digital text management – The SOLO categories," in press.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Siamese Swin Transformers for Robust Polyp Classification

Mesut ŞEKER
Department of Electrical-Electronics Engineering
Dicle University
Diyarbakır, Türkiye
mesut.seker@dicle.edu.tr
ORCID: 0000-0001-9245-6790

Abstract
Colorectal cancer prevention relies heavily on the accurate identification of precancerous polyps during screening programs. However, distinguishing between benign hyperplastic polyps and precancerous sessile serrated adenomas remains a significant clinical challenge due to complex morphological overlapping and high interobserver variability. Standard supervised deep learning models often struggle to generalize across these subtle inter-class differences, particularly when trained on noisy data and subject to persistent clinical label ambiguity. To address these limitations, we propose an end-to-end Joint Siamese Swin Transformer architecture tailored for automated stratification of colorectal polyps. The framework employs a multi-task hybrid objective function that synergizes a pairwise contrastive metric regularizer with a standard supervised classification loss. This geometric constraint guides the hierarchical attention backbone to distill highly discriminative, class-compact feature topologies, effectively mitigating the negative effects of label ambiguity. Comprehensive evaluations on the MHIST dataset demonstrate that our model achieves a superior overall accuracy of 86.49% and a macro F1 score of 85.50%, surpassing standard convolutional networks and vanilla transformer baselines. A rigorous subgroup analysis stratified by pathologist agreement levels reveals exceptional diagnostic resilience, with the model peaking at a 96.17% F1 score in the highest consensus tier. Furthermore, the architecture exhibits substantial structural efficiency, operating with only 19.51 million parameters and 2.98 GFLOPs, thereby reducing computational complexity by over 75% compared to conventional baselines. Ultimately, this study demonstrates that combining hierarchical vision transformers with Siamese metric learning yields a highly accurate, parameter-efficient, and edge-compatible solution for scalable digital pathology workflows.

Keywords*:* Computational Pathology, Colorectal Polyps, Swin Transformer, Contrastive Learning, Multi-task Optimization

I. INTRODUCTION

Accounting for approximately 10% of all global cases, colorectal cancer is the third most diagnosed malignancy and ranks as the second leading cause of cancer-related mortality worldwide. Although it predominantly affects individuals aged 50 and older, the incidence has been rising notably among younger adults aged 30 to 50, further emphasizing the critical need for robust diagnostic screening [1]. The foundation of preventing this disease relies heavily on screening programs that identify precancerous colorectal polyps. Consequently, histological examination of biopsy samples constitutes a significant portion of a pathologist's workload [2]. However, manual assessment of gigapixel whole-slide images is time-consuming, costly, and hampered by a global shortage of pathologists [3]. Furthermore, distinguishing between benign hyperplastic polyps and precancerous sessile serrated adenomas is hampered by significant interobserver variability due to complex morphological similarities. To address these clinical bottlenecks, researchers have increasingly integrated deep learning techniques into computational pathology to automate disease detection and staging[4].

Recent studies have demonstrated that deep learning architectures can achieve diagnostic accuracies comparable to those of medical experts. Standard convolutional neural networks and their modernized variants have been extensively tested in this domain. For instance, the ConvNeXt architecture was applied to the MHIST dataset and successfully classified precancerous lesions across various low data regimes and difficulty levels, outperforming traditional convolutional baselines [5]. To overcome the limitations of standard cross-entropy loss in convolutional networks, such as sensitivity to noise and poor margin boundaries, researchers proposed a Supervised Constructive Learning model. This model effectively combined a constructive loss function with standard supervision to improve classification accuracy across multiple colorectal cancer datasets [6]. Moreover, to reduce reliance on exhaustive manual annotations, self-supervised learning frameworks have gained traction. Specifically, optimized Barlow Twins architectures utilizing Swin Transformer encoders were developed to extract robust, invariant feature representations from unannotated pathology patches [7]. Despite these remarkable advancements, standard supervised networks continue to struggle to generalize across subtle inter-class differences and noisy data [8]. While self-supervised and constructive learning methodologies produce robust representations, current research lacks comprehensive methods that fully fuse metric contrastive learning with hierarchical Vision Transformers for end-to-end optimization [9]. It remains unknown whether standard attention mechanisms retain their diagnostic resilience when exposed to severe clinical label ambiguity. While other hybrid architectures exist, this study explores an alternative paradigm by specifically evaluating a tightly constrained joint-training framework designed for extreme annotator discordance, positioning computational efficiency alongside diagnostic resilience as co-primary objectives. Through this optimization, this study explicitly clarifies how joint geometric constraints influence structural cluster separation and practical classification boundaries under clinical ambiguity. We hypothesized that implementing a hybrid objective function that combines contrastive metric learning with supervised classification would enforce more discriminative topological boundaries, thereby increasing diagnostic accuracy in precancerous lesion stratification even under severe annotator discordance. We aimed to propose and evaluate an end-to-end Joint Training Siamese Swin Transformer architecture for colorectal polyp classification. Through a comprehensive evaluation of the MHIST dataset, we aimed to assess classification performance by measuring accuracy and macro F1 scores across four distinct clinical difficulty tiers defined by pathologist agreement levels.

II. MATERIAL AND METHODS

1. Material

In this study, the Minimalist Histopathology (MHIST) benchmark dataset was utilized to evaluate the proposed framework. The cohort consists of 3,152 microscopic colorectal polyp images with a uniform spatial resolution of 224 × 224 pixels. The dataset is pre-partitioned into a standardized split consisting of 2,175 images for training and 977 images for independent testing, ensuring that all benchmarking experiments follow a consistent evaluation protocol. Derived from clinical FFPE blocks and processed through standard H&E staining pipelines at the Dartmouth-Hitchcock Medical Center, the dataset is used for binary classification between Hyperplastic Polyps (HP) and Sessile Serrated Adenomas (SSA). Gold-standard annotations were adjudicated by a majority vote among seven independent expert pathologists. Inherent to its design, a notable fraction of the dataset features borderline diagnostic consensus (e.g., 4/7 agreement), introducing a non-trivial challenge that replicates real-world clinical subjectivity. The data collection and dissemination protocol received full clearance from the Dartmouth-Hitchcock Health Institutional Review Board[10].

In line with the literature, we further stratified the test partition into four subgroups based on pathologist agreement: Very Easy (7/7), Easy (6/7), Hard (5/7), and Very Hard (4/7) to systematically analyze model performance under human-annotator discordance [11].

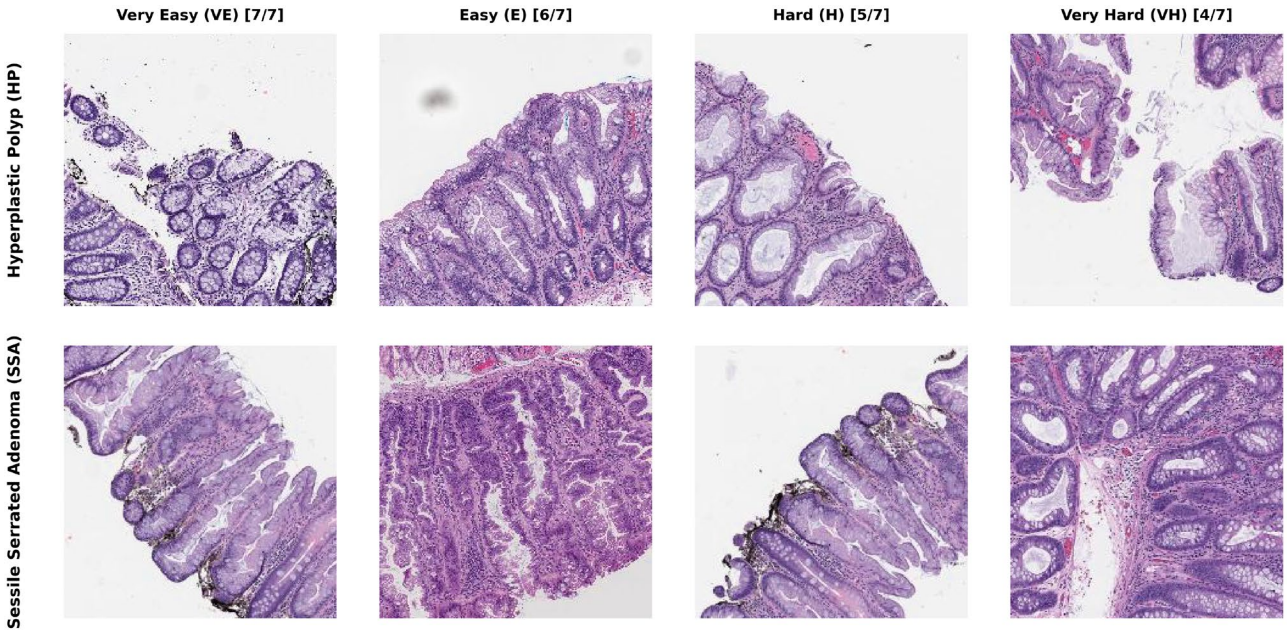


Fig. 1. Sample images of VE, E, H, and VD subgroups for SSA and HP classes

2. Method

The proposed study consists of seven main stages. First, histopathological images from the MHIST dataset were collected and divided into training and test sets. Second, image preprocessing and augmentation techniques, including resizing, normalization, flipping, rotation, and color jittering, were applied to enhance data diversity and improve model generalization. Third, positive and negative image pairs were generated for Siamese learning. Fourth, the paired images were processed using a shared Swin Transformer (Swin-T) backbone to extract discriminative features. Fifth, the extracted features were simultaneously fed into a projection head for contrastive learning and a classification head for HP–SSA classification. Sixth, the network was jointly optimized using a weighted combination of contrastive and cross-entropy losses. Finally, the trained model was evaluated using classification metrics, confusion matrix analysis, t-SNE visualization, and difficulty-level subgroup analysis. All these steps are shown in Fig. 1.

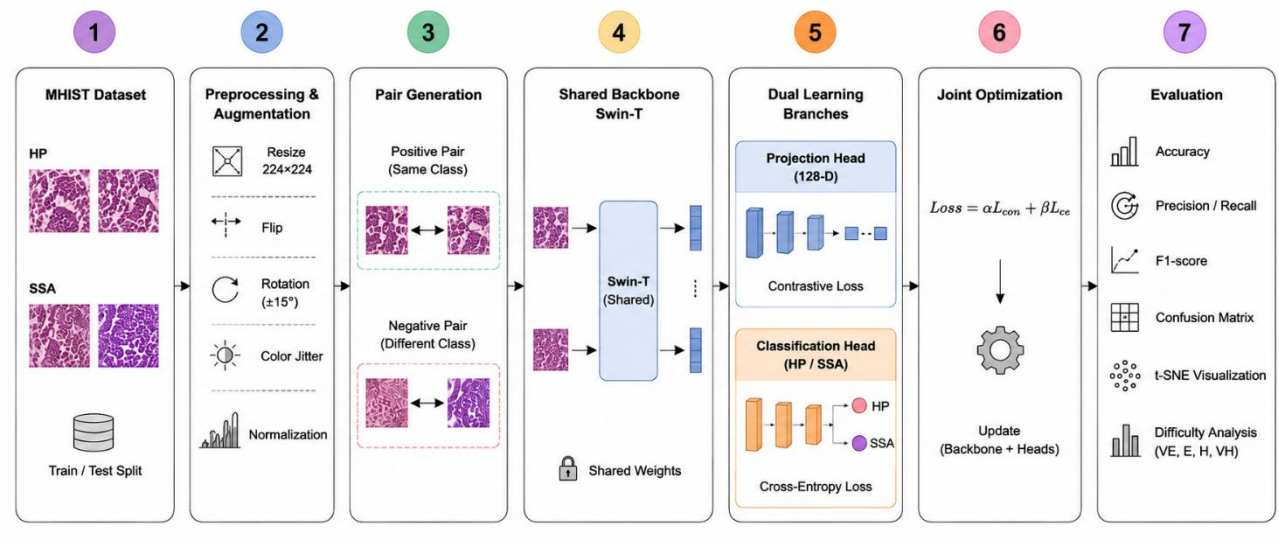


Fig. 2. Systematic workflow of the proposed study illustrating the Joint Siamese Swin Transformer pipeline for colorectal polyp stratification

3. Siamese Pair Construction

To enable metric learning, a Siamese training dataset is generated from the original training partition. For each anchor image, a second image from the same class was selected with 50% probability to form a positive pair; otherwise, an image from the opposite class was selected to form a negative pair. Accordingly, the target labels for contrastive learning were defined as 1 for positive pairs and 0 for negative pairs. This strategy allowed the network to learn discriminative relationships between histopathological samples.

4. Proposed Joint-Training Siamese Swin-T Architecture

The proposed framework employs an end-to-end Joint-Training Siamese Swin Transformer architecture, in which a shared Swin-T backbone, initialized with ImageNet-pretrained weights, extracts robust feature representations from paired images. To facilitate multi-task optimization, the original Swin-T classification head was removed and replaced with two parallel functional branches that directly originate from the backbone features. The first branch, designated the projection branch, was introduced to generate optimized embeddings for contrastive learning via a sequential module comprising a fully connected layer, a GELU activation function, and a second fully connected layer that produces a 128-dimensional embedding, which is subsequently L2-normalized before contrastive optimization. Simultaneously, a classification branch was integrated to perform direct polyp classification, consisting of a fully connected layer mapped from the rich backbone features, followed by batch normalization, a *ReLU* activation, dropout regularization, and a final fully connected layer that yields two class logits. Because both branches share the same feature extractor without layer freezing, the underlying backbone learns generalizable representations that benefit both discriminative classification and semantic similarity learning.

5. Joint Optimization Strategy

The proposed model was trained using a hybrid objective function combining contrastive learning and supervised classification. To minimize the distance between embeddings originating from the same class while simultaneously separating embeddings belonging to different classes, a contrastive loss function was employed to constrain the latent feature space geometrically. In parallel with this metric regularizer, a standard Cross-Entropy loss (L) was utilized within the classification branch to supervise the primary polyp categorization task by measuring the discrepancy between the predicted logits and the ground-truth labels. To jointly optimize both objective fields without feature degradation, a total hybrid loss function was formulated as a weighted linear combination of the contrastive and classification components, defined as $Ltotal = \alpha Lcon + \beta Lce$.In this multi-task optimization paradigm, the scaling hyperparameters were set to α = 0.1 and β = 1.0, thereby positioning the contrastive component as an auxiliary regularizer that guides the backbone toward learning class-compact topologies without overshadowing the primary classification task. The entire multi-task network was optimized end-to-end for 50 epochs using the *AdamW optimizer with a base learning rate of 1 x 10^-4 and a weight decay of 1 x 10^-4 to prevent overfitting under* data scarcity. Furthermore, to ensure stable convergence and fine-tune the final representation boundaries, a cosine annealing learning rate scheduler was integrated to smoothly decay the learning rate to a minimum threshold over the training duration.

6. Evaluation Stage

During the inference stage, the trained models were rigorously evaluated using the independent test partition. Crucially, the proposed Joint Siamese method required only a single image for testing, bypassing the need for image pairing during deployment. The shared Swin-T backbone generated target feature embeddings. At the same time, the parallel classification head simultaneously produced final binary diagnostic predictions, from which the learned latent representations were extracted for subsequent downstream visualization analyses. Model performance was comprehensively assessed using multiple complementary metrics and frameworks. The overall classification proficiency was quantified using standard accuracy and macro F1-scores, supplemented by empirical confusion matrices to analyze class-specific error distributions and prediction behavior. To further investigate the discriminative capability and structural topology of the learned representations, t-distributed Stochastic Neighbor Embedding (t-SNE) was employed to project the high-dimensional feature embeddings into a comprehensible two-dimensional manifold. Finally, a robust sub-group evaluation was conducted by reporting classification performance separately across the *VE, E, H, and VH* difficulty categories, thereby providing critical insights into model behavior and diagnostic resilience under varying levels of human annotator agreement.

7. Baseline Models

Two baseline classification architectures are investigated: ResNet-50 [12] and Swin Transformer – Tiny (SwinT) [13]. Both models were initialized using ImageNet-pretrained weights to leverage robust structural priors. Their original classification heads were replaced with a dedicated fully connected layer that produces two output logits for the HP and SSA classes. Unlike the proposed joint-training framework, these baseline models were trained independently using standard supervised learning, guided solely by a traditional Cross-Entropy loss function. They were optimized using the *AdamW* optimizer with identical hyperparameters to ensure a rigorous and fair benchmark comparison against our proposed multi-task architecture. To ensure full reproducibility, all experimental pipelines were implemented using Python 3.10 and PyTorch 2.1 on an Anaconda environment. The networks were trained on an enterprise-grade workstation equipped with an NVIDIA RTX 3060 GPU (12 GB VRAM) and an Intel Core i7 processor. Data augmentation including random horizontal/vertical flips,

rotation (±15º), and color jittering (brightness=0.2, contrast=0.2). To eliminate stochastic variance and ensure full reproducibility, all random number generators across the software ecosystem were seeded with 42.

III. RESULTS AND DISCUSSION

The experimental outcomes of the proposed Joint Siamese Swin Transformer, along with those of the baseline architectures, are systematically evaluated across overall diagnostic metrics, topological feature distributions, difficulty-stratified subgroups, and computational complexity.

1. Overall Classification Performance and Visual Analysis

The quantitative assessment of the proposed framework alongside baseline architectures is summarized in Table 1. The proposed network achieved superior global convergence across the test partition, achieving an overall accuracy of 86.49% and a macro F1-score of 85.50%. This performance establishes a distinct diagnostic threshold over both the standard Swin-T (83.73% accuracy, 82.15% F1-score) and ResNet-50 (84.85% accuracy, 82.89% F1-score) configurations.

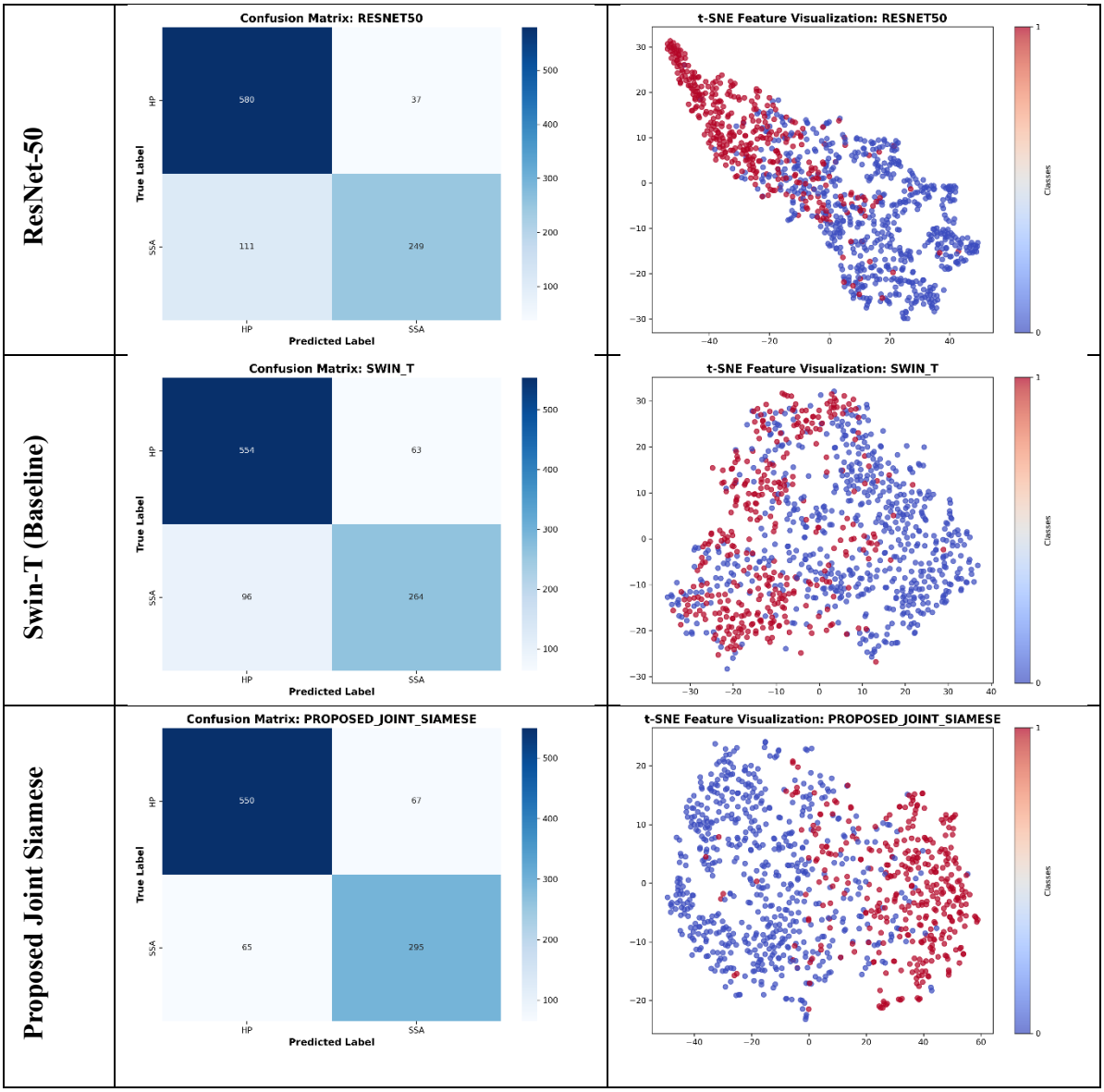
Table 1. Detailed performance breakdown and sub-group evaluation of the models on the MHIST test partition, stratified by the level of pathologist agreement. (VE: Very Easy with 7/7 consensus; E: Easy with 6/7 consensus; H: Hard with 5/7 consensus; VH: Very Hard with 4/7 consensus. Bold values indicate the highest performance achieved within each respective difficulty layer.

Difficulty Level	**Test Samples**	**ResNet-50 Acc / F1 (%)**	**Swin-T (Baseline) Acc / F1 (%)**	**Proposed Joint Siamese Acc / F1 (%)**
Overall	977	84.85 / 82.89	83.73 / 82.15	86.49 / 85.50
Very Easy (VE) [7/7]	323	93.50 / 93.22	95.05 / 94.95	96.28 / 96.17
Easy (E) [6/7]	304	90.46 / 88.12	91.12 / 89.97	93.42 / 92.34
Hard (H) [5/7]	186	76.34 / 70.28	78.49 / 76.52	80.65 / 78.40
Very Hard (VH) [4/7]	164	68.90 / 64.63	69.51 / 69.14	64.02 / 63.42

This numerical superiority is comprehensively corroborated by the visual feature space analysis presented in Table 2. The t-SNE projections explicitly reveal that the proposed Joint Siamese network constructs more compact intra-class topologies and wider inter-class margins compared to the scattered latent representations of the baseline models. Consequently, the empirical error distributions in the confusion matrices show more balanced and precise classification between the Hyperplastic Polyp and Sessile Serrated Adenoma classes, thereby reducing false-positive rates.

Table 2. Overall confusion matrices and t-SNE visualizations of the evaluated models on the MHIST test partition (N=977)

Models	**Confusion Matrices**	**t-SNE**



2. Sub-group Evaluation and Diagnostic Resilience

As detailed in Table 1, a rigorous subgroup evaluation was conducted to assess model resilience across varying levels of diagnostic difficulty and expert subjectivity. Analysis stratified by pathologist agreement levels revealed that the proposed framework consistently maintains its edge in clinically clear cases. The architecture achieved an outstanding 96.17% F1-score in the Very Easy (7/7 consensus) partition and maintained high resilience across the Easy (6/7 consensus) and Hard (5/7 consensus) subsets, significantly outperforming the vanilla transformer model.

A critical inspection of the subgroup analysis in Table 1 reveals a localized performance inversion in the Very Hard (VH) cohort [4/7 agreement]. In this extreme discordance tier, the proposed Joint Siamese network underperforms, achieving an accuracy of 64.02% compared to the baseline Swin-T's 69.51%. This drop

highlights a core limitation of rigid metric learning constraints when applied to clinical ambiguity. The pairwise contrastive loss enforces a strict geometric distance (margin = 1.0), pushing different classes apart and pulling identical classes together based on the majority-vote label. However, a 4/7 agreement indicates that even human experts are nearly evenly split, reflecting significant biological and morphological overlap (e.g., borderline serrated features). The forces of the contrastive objective, by forcing these fundamentally ambiguous transitionary features into rigid, distinct geometric clusters, over-penalize the latent representations. While standard cross-entropy in the vanilla Swin-T retains a degree of probabilistic flexibility that tolerates label noise, our joint geometric constraints distort borderline feature topologies under severe clinical ambiguity. This limitation underscores the fact that while Siamese regularization yields superior class-compact representations for distinct phenotypes (VE and E), it compromises performance at the extreme, non-consensus boundaries of digital pathology workflows.

3. Computational Complexity and Efficiency

Strikingly, the diagnostic accuracy accomplished by the proposed network is accompanied by substantial computational savings, as explicitly outlined in Table 3. The proposed Joint Siamese framework operates with just 19.51M parameters and a mere 2.98 GFLOPs. This architectural optimization reduces computational complexity by approximately 75% relative to the baseline Swin-T (which requires 37.70 M parameters and 11.91 GFLOPs) and by over 80% compared to the computationally intensive ResNet-50 model (47.02 M parameters and 16.53 GFLOPs). These findings demonstrate that the proposed multi-task architecture achieves state-of-the-art diagnostic performance while maintaining a lightweight footprint, making it well-suited for scalable, real-world digital pathology applications and edge computing workflows.

Table 3. Computational complexity, parameter counts, and GFLOPs comparison of the evaluated models

Models	**Params (M)**	**GFLOPs**
ResNet-50	47.02	16.53
Swin-T (Baseline)	37.70	11.91
Proposed Joint Siamese	19.51	2.98

4. Benchmarking Against Recent Studies

As summarized in Table 4, the comparative landscape of recent methodologies applied to the MHIST dataset reveals a persistent tradeoff between computational efficiency and diagnostic resilience. While modernized convolutional networks and heavily parameterized hybrid architectures achieve notable accuracy, they often require substantial computational resources or rely on overly complex pretraining paradigms that hinder real-world clinical deployment. Conversely, broad foundation models and standard self-supervised frameworks often lack the specialized fine-tuning required to navigate the subtle morphological overlaps among colorectal polyps. Positioned within this context, the proposed Joint Siamese Swin Transformer bridges a critical gap in the literature. By infusing hierarchical attention mechanisms with multi-task contrastive learning, the proposed framework delivers highly competitive diagnostic performance while operating with a fraction of the

computational footprint of existing models. This strategic optimization not only mitigates the noise sensitivity typical of standard supervised networks but also establishes a scalable, edge-compatible solution for automated digital pathology.

When benchmarked against recently published methodologies, it is critical to contextualize absolute accuracy with respect to underlying computational complexity and architectural requirements. For instance, while the model proposed by Fırıldak et al. [6] reports near-perfect accuracy of 99.52%, this performance is achieved through specialized multi-module configurations (RFB, DWC, and GAP) that introduce architectural complexity. Conversely, our approach intentionally prioritizes structural minimalism and parameter efficiency. Achieving an overall accuracy of 86.49% and an F1-score of 85.50% represents a highly competitive trade-off, given that our framework operates with only 19.51M parameters and 2.98 GFLOPs, which translates to an approximate 75% reduction in computational overhead compared to vanilla transformers. In clinical pathology workflows where edge-computing and low-latency throughput are paramount, absolute accuracy cannot be evaluated in isolation from hardware footprints. Our framework is therefore preferable not for achieving an absolute peak in accuracy, but for providing an optimized, scalable balance for resource-constrained clinical deployment.

Table 4. Comparison of State-of-the-Art Studies on the MHIST Dataset

Author (Study)	**Method**	**Objective**	**Results (MHIST Only)**	**Limitations / Gaps**
Notton et al. (2026) [7]	Self-Supervised Barlow Twins (ResNet-50 and Swin-T)	To optimize the Barlow Twins framework for histopathology data and investigate the optimal Field of View for colorectal polyp screening.	Accuracy: 84.95% (basicBT) AUC: 82.64% (imBT).	The proposed augmentation strategies did not generalize well to downstream tasks such as MHIST, and weak color jitter reduced the model's adaptability to other stains.
Nergiz (2023) [5]	ConvNeXt (Modernized ResNet-50)	To evaluate the classification of precancerous colorectal lesions on the MHIST dataset across full data, k-shot, and gradually increasing difficulty scenarios.	Accuracy: 88.90% AUC: 93.91% F1-Score: 0.9121.	Requires further validation on larger, more comprehensive datasets better to analyze the model's behavior in complex topologies.
Firildak et al. (2025) [6]	Supervised Constructive Learning (SCL) with RFB, DWC, and GAP modules	To achieve highly accurate colorectal cancer detection by overcoming the disadvantages of standard CNN cross-entropy loss, such as poor margin	Accuracy: 99.52% AUC: 99.49% F1-Score: 99.40%.	Under conditions of extreme annotator discordance, the strict clustering boundaries established by the contrastive approach may cause performance fluctuations.

		possibility and noise sensitivity.		
Wang et al. (2021) [14]	CNN + Transformer + Token-aggregating and excitation (TAE)	To build a hybrid classification architecture on histopathological images using self-supervised learning.	Accuracy: 83.02% AUC: 94.76% F1-Score: 85.86%	It is impractical for small research groups as it requires millions of histopathology images, 32 GPUs, and 200 hours of training time.
Wei et al. (2020) [15]	Curriculum Learning / ResNet-based	To achieve model calibration by applying a training strategy with increasing levels of difficulty based on pathologist agreement.	AUC: 93.6%.	It is overly dependent on detailed pathologist consensus labels to determine data difficulty levels during training.
Springenberg et al. (2023) [16]	ConvNeXt-L	To compare the classification robustness of modern CNN architectures and Vision Transformers on histopathology datasets.	Accuracy: 88.56% AUC: 95.39%.	As a primarily comparative analysis, it does not propose a novel end-to-end diagnostic architecture or a metric loss specific to colorectal tissues.
Kaiko.ai et al. (2024) [17]	DINO ViT-S8 (Foundation Model)	To conduct large-scale and multi-source training of pathology foundation models.	Accuracy: 83.2%.	The classification performance remains low compared to specialized architectures fine-tuned directly on the MHIST dataset.
This study	Joint Siamese Swin Transformer	To overcome morphological ambiguity and inter-observer variability using a parameter-efficient, multi-task architecture	Accuracy: 86.49% F1-Score: 85.50%	Contrastive loss constraints face minor limitations under extreme annotator discordance (Very Hard) due to severe label noise.

5. Architectural Design Decisions and Parameter Rationale

The configuration of the proposed multi-task architecture relies on distinct design choices that have been validated through empirical stabilization. First, the selection of the Swin Transformer backbone over standard CNNs was dictated by its hierarchical attention mechanism, which naturally captures multi-scale tissue morphologies essential for digital pathology. Second, regarding the contrastive loss weighting ($\boldsymbol{\alpha = 0.1}$),

fitting this parameter as an auxiliary regularizer was essential; preliminary calibrations indicated that larger weights over-penalized the latent space, forcing borderline samples into premature clusters and degrading downstream classification. Finally, our augmentation strategy (incorporating rotation and color jittering) was specifically tailored to replicate the staining variations and physical tissue rotations common in multi-center clinical blocks, thereby ensuring that the learned Siamese topologies are robust to non-biological noise rather than memorizing batch artifacts.

IV. CONCLUSION

This study presents an end-to-end Joint Siamese Swin Transformer architecture to overcome the frequent morphological ambiguities and high observer variability in colorectal polyp stratification. Unlike conventional approaches in literature, this multi-task hybrid architecture does not rely solely on standard classification supervision. Still, it incorporates a contrastive topological constraint, successfully striking a unique balance between parameter efficiency and diagnostic performance. The proposed model demonstrated superior generalization, particularly in cases with high pathologist consensus, while drastically reducing hardware overhead compared to standard deep learning models.

However, the proposed methodology has certain limitations. In highly challenging cases characterized by extreme clinical label noise and minimal pathologist agreement, the rigid geometric constraints imposed by contrastive optimization slightly reduced the network's decision flexibility. Furthermore, the current architectural evaluation was restricted to a specific binary classification task.

In future studies, the current architecture will be adapted for multi-class and more extensive histopathological datasets to advance its clinical applicability further. Additionally, to enhance model resilience in borderline tissue samples with high label ambiguity and to address data scarcity, we plan to integrate advanced generative artificial intelligence-based data augmentation strategies, such as conditional diffusion or structure-preserving style transfer, into this diagnostic pipeline.

ACKNOWLEDGEMENT

This study has been supported by Dicle University Scientific Research Projects (DÜBAP) Coordinatorship with the Project Number MÜHENDİSLİK.26.029.

REFERENCES

[1] "Colorectal cancer," WHO. Accessed: Jun. 22, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer

[2] B. A. Sullivan, M. Noujaim, and J. Roper, "Cause, Epidemiology, and Histology of Polyps and Pathways to Colorectal Cancer," *Gastrointest. Endosc. Clin. N. Am.*, vol. 32, no. 2, pp. 177–194, Apr. 2022, doi: 10.1016/j.giec.2021.12.001.

[3] S. A. Omoush, J. A. M. Alzyoud, N. K. T. El-Omari, and A. J. A. Alzyoud, "The Role of Whole Slide Imaging in AI-Based Digital Pathology: Current Challenges and Future Directions—An Updated Literature Review," *J. Mol. Pathol.*, vol. 7, no. 1, p. 2, Jan. 2026, doi: 10.3390/jmp7010002.

[4] D. Seker, A. Akhan, and A. Yıldız, "A Deep-XAI Method for Histopathological Image Classification: Utilizing Transformer-based FNet Architecture and LIME Method," *Pamukkale Univ. J. Eng. Sci.*, vol. 31, no. 6, 2025, doi: 10.5505/pajes.2025.98572.

[5] M. NERGİZ, "Classification of Precancerous Colorectal Lesions via ConvNeXt on Histopathological

Images," *Balk. J. Electr. Comput. Eng.*, vol. 11, no. 2, pp. 129–137, 2023, doi: 10.17694/bajece.1240284.

[6] K. Firildak, G. Celik, and M. F. Talu, "Supervised Constructive Learning-Based Model for Identifying Colorectal Cancer Tissue Types From Histopathological Images," *Int. J. Imaging Syst. Technol.*, vol. 35, no. 4, 2025, doi: 10.1002/ima.70161.

[7] C. Notton *et al.*, "Efficient self-supervised Barlow Twins from limited tissue slide cohorts for colonic pathology diagnostics," *Med. Image Anal.*, vol. 111, no. November 2024, 2026, doi: 10.1016/j.media.2026.104004.

[8] M. Liu, "RAUM-Net: Regional Attention and Uncertainty-aware Mamba Network," Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.21905

[9] A. Waseem, P. Ruiu, S. Nixon, A. Lagorio, and M. Tistarelli, "Self-attention as the backbone: A survey on Vision Transformers," *Comput. Vis. Image Underst.*, vol. 269, p. 104791, 2026, doi: https://doi.org/10.1016/j.cviu.2026.104791.

[10] J. Wei *et al.*, "A Petri Dish for Histopathology Image Analysis BT - Artificial Intelligence in Medicine," A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, and D. Riaño, Eds., Cham: Springer International Publishing, 2021, pp. 11–24.

[11] N. A. C. S. Wong, L. P. Hunt, M. R. Novelli, N. A. Shepherd, and B. F. Warren, "Observer agreement in the diagnosis of serrated polyps of the large bowel," *Histopathology*, vol. 55, no. 1, pp. 63–66, Jul. 2009, doi: 10.1111/j.1365-2559.2009.03329.x.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[13] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2103.14030

[14] X. Wang *et al.*, "TransPath: Transformer-Based Self-supervised Learning for Histopathological Image Classification," 2021, pp. 186–195. doi: 10.1007/978-3-030-87237-3_18.

[15] J. Wei *et al.*, "Learn like a Pathologist: Curriculum Learning by Annotator Agreement for Histopathology Image Classification," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.13698

[16] M. Springenberg, A. Frommholz, M. Wenzel, E. Weicken, J. Ma, and N. Strodthoff, "From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology," *Med. Image Anal.*, vol. 87, p. 102809, 2023, doi: https://doi.org/10.1016/j.media.2023.102809.

[17] kaiko. Ai *et al.*, "Towards Large-Scale Training of Pathology Foundation Models," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2404.15217

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Architecture of a Fuzzy Logic Model for Supporting Investigations in Burglary Cases

Radmila Radostinova Roydeva
Department of Computer Systems and Technologies
Burgas Free University
Burgas, Bulgaria
radmila.roydeva@bfu.bg
ORCID: 0009-0005-0064-7765

Penka Georgieva
Department of Computer Systems and Technologies
Burgas Free University
Burgas, Bulgaria
penka.georgieva@bfu.bg
ORCID: 0000-0001-9598-5997

Abstract
Criminal investigations are characterized by uncertainty, incomplete information, dynamic evidence, and the simultaneous existence of multiple competing investigative hypotheses. Traditional analytical approaches often rely heavily on subjective expert judgment, which can lead to inconsistencies in interpreting and prioritizing investigative versions. This paper presents the conceptual design and architecture of a fuzzy-logic–based decision-support model intended to facilitate the generation of investigative hypotheses in burglary cases. The proposed system is implemented as a Mamdani-type fuzzy inference model, designed to formalize expert investigative reasoning using linguistic variables and rule-based inference. The architecture incorporates twelve input variables derived from forensic, behavioral, and criminological indicators commonly encountered in burglary investigations. These variables are transformed into linguistic representations using fuzzy membership functions and processed through an expert-defined rule base. The system produces five ranked investigative hypotheses corresponding to distinct investigative scenarios, including serial criminal activity, opportunistic offenses, coordinated participation, alternative explanations, and possible staged events related to insurance motives. The paper provides a detailed explanation of the architectural design process, including the selection of input and output variables, the construction of linguistic terms, the development of membership functions, and the rationale for the adopted inference mechanism. Particular attention is given to the model's interpretability and its suitability for supporting analytical reasoning under conditions of uncertainty. The proposed architecture demonstrates how fuzzy logic can model complex investigative reasoning processes while maintaining transparency and expert interpretability. The study establishes the methodological foundation for future experimental validation and further development of intelligent forensic decision support systems.

Keywords: Fuzzy logic, investigative analysis, burglary investigation, decision support systems, forensic reasoning, fuzzy inference systems, criminology

I. INTRODUCTION

The process of criminal investigation involves the continuous analysis and interpretation of incomplete, uncertain, and often contradictory information. Investigators are required to formulate and evaluate multiple competing hypotheses while working under conditions of limited evidence, dynamic situational changes, and varying levels of information reliability. This problem is particularly significant in burglary investigations, where the same factual circumstances may support several alternative investigative interpretations. The construction and evaluation of investigative hypotheses are fundamental to criminal investigation and analytical reasoning. The investigative version serves as a methodological instrument for explaining available facts and directing further evidentiary activities [1]. One of the fundamental challenges in investigative analysis is that many criminological indicators cannot be interpreted in a purely binary manner. Indicators such as the degree

of force used during entry, the level of offender preparation, the existence of coordination between participants, or the possibility of an alternative explanation are inherently gradual rather than absolute. In practice, investigators rarely classify a circumstance as completely present or completely absent. Instead, such assessments are usually expressed linguistically as terms like "low," "moderate," or "high." This characteristic makes classical binary logic insufficient for modeling real investigative reasoning processes. Fuzzy logic was first introduced by Zadeh in 1965 and has since become one of the most widely applied approaches for modeling uncertainty and approximate reasoning [2]. The theoretical foundations and practical applications of fuzzy systems have been extensively described in the literature [3], [4]. This study presents the conceptual design and architecture of a fuzzy-logic–based decision-support model for generating investigative hypotheses in burglary cases. The model was implemented using MATLAB Fuzzy Logic Toolbox [5]. The Mamdani approach remains one of the most frequently used fuzzy inference methods due to its interpretability and suitability for expert systems [6]. The architecture integrates forensic, behavioral, and criminological indicators into a unified analytical framework that generates ranked investigative versions. Recent research highlights the growing role of artificial intelligence in digital forensic analysis and the importance of explainability in legal applications [7], [8]. Unlike predominantly data-driven AI approaches, the proposed model uses an explicit fuzzy rule base, making the relationship between investigative indicators and generated hypotheses transparent and traceable.

II. CONCEPTUAL ARCHITECTURE

The proposed system is a fuzzy-logic–based decision-support architecture intended to assist the analytical process of generating investigative hypotheses in burglary cases, following the broader principles of artificial intelligence and knowledge-based decision-support systems [9], [10]. The model follows the structure of a Mamdani-type fuzzy inference system and consists of several sequential stages: acquisition of investigative indicators, fuzzification of input variables, rule-based inference, aggregation of fuzzy conclusions, defuzzification, and generation of ranked investigative hypotheses.

The conceptual workflow of the system can be summarized as follows:
Investigative facts → Input indicators → Fuzzification →
Rule evaluation → Aggregation → Defuzzification →
Ranked investigative hypotheses

The architecture is designed to imitate the reasoning logic of criminal investigators operating under uncertainty. Instead of applying strict deterministic conditions, the system allows simultaneous partial activation of multiple investigative hypotheses. This feature reflects the realistic conditions of criminal investigations, where available evidence rarely supports a single interpretation with absolute certainty.
The system uses 12 input variables representing criminological, behavioral, forensic, and situational indicators commonly encountered in burglary investigations. Each variable is evaluated over a normalized interval from 0 to 10 and mapped to fuzzy linguistic values. The fuzzy inference engine processes these variables through an expert-defined rule base. The activated rules produce fuzzy output values corresponding to five distinct investigative versions. The final stage applies centroid defuzzification to transform aggregated fuzzy outputs into numerical scores in the [0–100] interval, allowing the hypotheses to be ranked by their degree of support.
The overall architecture emphasizes interpretability and transparency. Every stage of the reasoning process remains traceable and understandable, which is particularly important in legal and forensic contexts where analytical explainability is essential. Figure 1 presents the overall conceptual architecture of the proposed fuzzy-logic–based decision support system for generating investigative hypotheses in burglary cases. The model follows the structure of a Mamdani-type fuzzy inference system and is organized into several sequential analytical stages.

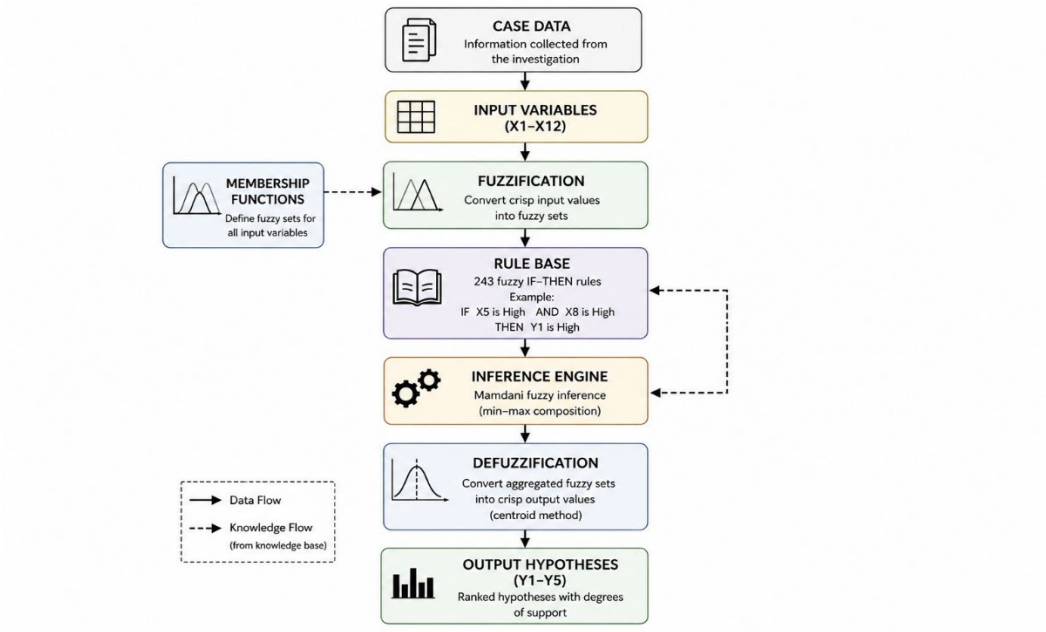


Figure 1. Architecture of the Proposed Fuzzy Inference System

The process begins with the collection and interpretation of investigative case data, including forensic traces, behavioral indicators, witness information, and contextual evidence obtained during the investigation. These data are transformed into twelve normalized input variables (X1–X12), each representing a specific criminological or forensic characteristic relevant to burglary analysis. During the fuzzification stage, the numerical input values are converted into linguistic representations using predefined membership functions. Each variable is described through three linguistic categories: low, medium, and high, which enable the model to represent uncertainty and gradual transitions between investigative conditions. The fuzzified inputs are processed by an expert-defined rule base comprising fuzzy IF–THEN rules derived from criminological reasoning and judicial case analysis. The rules model the logical relationships between investigative indicators and possible investigative hypotheses. The inference engine evaluates the activated rules using Mamdani min–max composition and aggregates the resulting fuzzy conclusions. The aggregated outputs are subsequently transformed into crisp numerical values through centroid defuzzification. The final stage generates five ranked investigative hypotheses (Y1–Y5), corresponding to different investigative interpretations of the burglary event. These outputs represent degrees of analytical support rather than deterministic conclusions and are intended to assist investigators during the hypothesis-generation process. The architecture emphasizes transparency, interpretability, and expert-oriented reasoning, making it suitable for forensic and legal analytical environments where explainability is essential.

III. SELECTION OF INPUT VARIABLES

The selection of input variables is one of the most critical stages in designing the proposed fuzzy inference model. The variables were chosen based on criminological theory, forensic investigative practice, judicial case analysis, and the need to capture indicators commonly associated with burglary investigations. The selection of investigative indicators was influenced by established criminalistics theory and investigative practice [11], [12].

The model incorporates twelve input variables.

Table 1. Input Variables

Variable	Input Indicator	Investigative Significance
X1	Degree of Forced Entry	Indicates the level of physical force, offender preparation, and possible professional criminal behavior during unauthorized entry.
X2	Use of Technical Tools	Reflects the use of specialized burglary instruments and technical means, suggesting planning and offender expertise.
X3	Displacement or Preparation of Objects	It represents the rearrangement, preparation, or selective handling of objects within the crime scene.
X4	Scope and Value of Missing Property	Evaluates the quantity and significance of stolen property.
X5	Seriality of Criminal Activity	Indicates repeated or patterned criminal behavior.
X6	Coordination or Conspiracy	Reflects indications of collaboration or coordinated actions.
X7	Disposal or Distribution of Stolen Property	Represents evidence regarding the handling and transfer of stolen items.
X8	Forensic Identification Evidence	Measures the strength of forensic identification evidence.
X9	Previous Criminal Activity	Indicates prior criminal behavior or recidivism.
X10	Possibility of Alternative Explanation	Reflects the probability of a lawful or non-criminal explanation.
X11	Insurance-Related Interest	Represents possible financial motivation related to insurance claims or fraud.
X12	Knowledge of the Target Object	Indicates offender familiarity with the target property.

The selected input variables were designed to capture the most significant forensic, behavioral, and criminological characteristics associated with burglary investigations. The variables combine objective forensic indicators with contextual investigative assessments to model the uncertainty inherent in real investigative environments. Particular attention was given to variables associated with offender preparation, coordination, behavioral repetition, and alternative explanatory scenarios. This combination enables the proposed fuzzy inference system to evaluate not only direct indicators of criminal activity but also contextual factors that may influence the interpretation of investigative evidence. The inclusion of variables such as insurance-related interests and alternative explanations is especially important because it enables the system to maintain investigative neutrality and reduce the risk of premature conclusions during the analytical phase.

IV. OUTPUT - INVESTIGATIVE HYPOTHESES

The proposed model generates five output hypotheses, each representing an alternative investigative direction.

Y1 – Serial or Professional Burglary

This hypothesis corresponds to organized, repeated, or technically sophisticated criminal activity.

Y2 – Opportunistic or Isolated Burglary

This version reflects spontaneous or low-planning criminal behavior committed without significant organization.

Y3 – Coordinated Participation or Conspiracy

This hypothesis addresses situations involving multiple offenders, collaboration, or coordinated criminal activity.

Y4 – Alternative Explanation or Insufficient Evidence

This version captures cases in which available information does not sufficiently support a finding of criminal activity or in which alternative explanations remain plausible.

Y5 – Staged Event or Insurance-Related Motivation

This hypothesis applies to situations involving the possible simulation of criminal activity for financial or insurance-related gain.

The selection of these five hypotheses is intended to cover the primary investigative directions encountered in burglary investigations while maintaining model interpretability and manageable rule complexity.

The proposed system generates five alternative investigative hypotheses representing the primary analytical directions commonly encountered in burglary investigations. Rather than producing deterministic classifications, the model assigns degrees of support to each hypothesis based on the evaluated investigative indicators.

V. LINGUISTIC VARIABLES AND MEMBERSHIP FUNCTIONS

The proposed system uses linguistic modeling to represent investigative uncertainty. Each input and output variable is described using three linguistic terms:

- Low
- Medium
- High

Figure 2 illustrates the fuzzy membership functions used for the linguistic representation of the input variables and output hypotheses within the proposed fuzzy inference system. The figure demonstrates how numerical investigative indicators are transformed into gradual linguistic assessments through overlapping fuzzy sets.

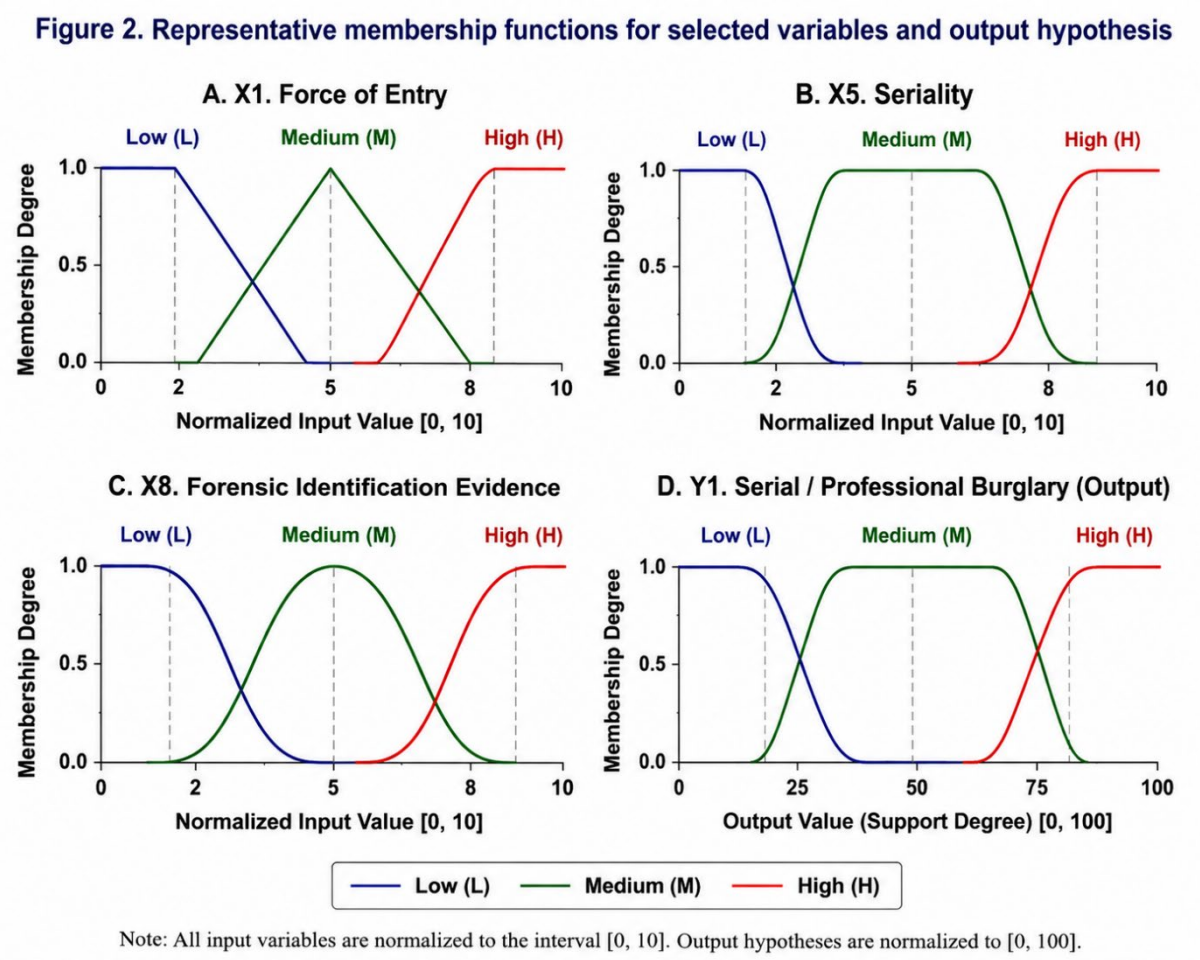


Figure 2. Representative membership functions for selected input variables and output hypotheses.

Different membership function shapes were intentionally selected based on the semantic meaning of the corresponding variable. Trapezoidal functions are primarily used for extreme linguistic categories ("Low" and "High") because they provide a stable representation near boundary values and allow stronger activation of clearly expressed conditions. Triangular functions are mainly used for intermediate ("Medium") states because they provide smooth transitions between neighboring categories.

Several input variables exhibit asymmetric or modified membership distributions. For example, the membership functions for variables associated with serial criminal behavior, alternative explanations, and insurance-related interests were adjusted to reflect the non-linear nature of investigative interpretation. Such asymmetry allows the model to represent situations better in which certain investigative indicators become analytically significant only after exceeding specific thresholds.

VI. RULE-BASED DESIGN

The rule base represents the core analytical mechanism of the proposed fuzzy inference system. The rules were developed through analysis of real judicial cases, criminological reasoning, and expert investigative logic.

Each rule follows the structure:

IF investigative conditions are satisfied

Then, the corresponding investigative hypothesis is activated

The model contains multiple groups of rules associated with different investigative hypotheses. Certain rules represent highly specific criminal patterns, while others serve as generalized fallback rules designed to ensure broader case coverage.
The use of weighted rules allows stronger indicators to exert greater influence on final outputs. This improves the model's flexibility and enables gradual refinement during optimization.
Importantly, the rules do not determine guilt or legal responsibility. Their role is strictly analytical and supportive, assisting investigators in structuring and prioritizing investigative hypotheses.

The model is implemented as a Mamdani-type fuzzy inference system. The Mamdani approach was selected for its interpretability, intuitive structure, and suitability for expert-defined reasoning.
The inference process consists of the following stages:

- Fuzzification of input values
- Rule activation
- Aggregation of fuzzy outputs
- Defuzzification of aggregated results

The model uses the minimum operator for conjunction, the maximum operator for disjunction, output aggregation, and centroid defuzzification for final numerical output generation.
Centroid defuzzification was selected because it provides balanced and stable numerical representations of aggregated fuzzy outputs. The method is widely accepted in fuzzy control systems and is particularly suitable when multiple rules contribute simultaneously to the final result.

VII. MATHEMATICAL REPRESENTATION OF THE PROPOSED FUZZY INFERENCE MODEL

The proposed investigative decision support system is mathematically modeled as a Mamdani-type fuzzy inference system comprising input variables, fuzzy membership functions, inference rules, and defuzzification procedures. Let the fuzzy inference system be represented as:

$$BüyükBüyükFIS = (X, Y, F, R)$$where:

$X = \{X_1, X_2, \ldots, X_{12}\}$represents the set of input variables;
$Y = \{Y1, Y2, ..., Y5\}$ represents the set of output investigative hypotheses;
$represents$ the set of fuzzy membership functions;
R represents the rule base.

1. Input Space

Each input variable is normalized within the interval:

$$X_i \in [0,10]$$

where the numerical values correspond to the degree of presence of a specific investigative indicator.
The output hypotheses are represented within the interval:

$$Y_i \in [0,100]$$

which enables finer differentiation between competing investigative versions.

2. Fuzzification Process

The fuzzification stage converts crisp numerical input values into fuzzy linguistic values using membership functions. In the proposed model, several types of membership functions are used, depending on the variable's analytical meaning. Triangular and trapezoidal membership functions are used to delineate low,

medium, and high linguistic categories clearly. In contrast, smoother, nonlinear functions, including S-, Z-, and bell-shaped forms, are used for variables for which gradual transitions better reflect the indicator's investigative interpretation. A fuzzy membership function is defined as:

$$\mu_A(x): X \to [0,1]$$

where:

X is the crisp input value;

$\mu_A(x)$ represents the degree of membership of X in fuzzy set A.

Each variable is represented through three linguistic categories:

Low (L), Medium (M), High (H)

Triangular membership functions are defined as:

$$\mu_A(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a < x \le b \\ \dfrac{c-x}{c-b}, & b < x < c \\ 0, & x \ge c \end{cases}$$

Trapezoidal membership functions are defined as:

$$\mu_A(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a < x \le b \\ 1, & b < x \le c \\ \dfrac{d-x}{d-c}, & c < x < d \\ 0, & x \ge d \end{cases}$$

The overlapping structure of the membership functions enables smooth transitions between neighboring linguistic categories and allows simultaneous partial activation of multiple fuzzy sets. The mathematical properties of fuzzy sets, membership functions, and computational intelligence methods are discussed extensively in the literature [2], [6], [8], [9].

3. Fuzzy Rule Base

The system's knowledge base consists of expert-defined fuzzy IF–THEN rules.
A general fuzzy rule can be represented as:

IF X1 is High AND X5 is High AND X9 is High, THEN Y1 is High

Formally, a fuzzy rule may be expressed as:

$$R_k: \left(X_1 \in A_1^k\right) \wedge \left(X_2 \in A_2^k\right) \wedge \cdots \wedge \left(X_n \in A_n^k\right) \to \left(Y \in B^k\right)$$

Neuro-fuzzy and hybrid fuzzy approaches have been investigated extensively in intelligent systems research [3].

4. Mamdani Inference Mechanism

The proposed architecture applies Mamdani min–max inference.
The activation level of a rule is computed using the minimum operator:

$$\alpha_k = \min\left(\mu_{A_1}(x_1), \mu_{A_2}(x_2), \dots, \mu_{A_n}(x_n)\right)$$

The outputs of all activated rules are aggregated using the maximum operator:

$$\mu_{agg}(y) = \max(\mu_1(y), \mu_2(y), \dots, \mu_n(y))$$

This mechanism enables simultaneous support for multiple investigative hypotheses.

5. Defuzzification

The aggregated fuzzy outputs are fuzzified into crisp numerical values using the centroid method.
The centroid method is defined as:

$$y^* = \frac{\int y\mu_{agg}(y)dy}{\int \mu_{agg}(y)dy}$$

Centroid defuzzification was selected because it provides balanced numerical estimates and stable analytical behavior in the presence of multiple rules simultaneously activated.

VIII. RESULTS AND DISCUSSION

A preliminary case-based evaluation of the proposed fuzzy inference architecture was conducted using 15 real judicial cases related to burglary and associated investigative scenarios. The cases were analyzed to identify recurring factual, forensic, behavioral, and situational patterns and to examine their correspondence with the twelve input variables and five output investigative hypotheses defined in the model. The selected cases covered heterogeneous patterns, including serial or professional burglary, isolated or opportunistic offending, coordinated participation, alternative explanations, and situations requiring verification of possible staging or insurance-related interest. The analysis resulted in a rule base comprising 36 fuzzy IF–THEN rules, structured around the five output hypotheses Y1–Y5. The rules encode recurring relationships between investigative indicators and plausible investigative interpretations. Representative rules include: IF Force of Entry is High AND Modus Operandi Similarity is High AND Seriality is High, THEN Serial/Professional Burglary (Y1) is High; IF Offender Preparation is Low AND Seriality is Low, THEN Isolated/Opportunistic Burglary (Y2) is High; and IF Coordination Indicators are High, THEN Coordinated Participation (Y3) is High. The rule structure also includes combinations supporting alternative explanations (Y4) and the verification of possible staged or insurance-related events (Y5). The case-based evaluation demonstrated that the proposed architecture can represent heterogeneous investigative patterns within a unified reasoning framework. Cases characterized by repeated modus operandi, offender preparation, technical means, and recidivism primarily supported Y1, whereas lower levels of preparation and seriality were associated with Y2. Evidence of coordinated actions increased support for Y3, while evidentiary inconsistencies or insufficient indications of unlawful appropriation

activated Y4. Indicators associated with possible staging or insurance-related interest contributed to Y5. These results demonstrate the logical consistency and interpretability of the rule base across different investigative scenarios. The present evaluation should be regarded as preliminary. The current study focuses on case-based calibration, logical consistency, and interpretability rather than large-scale predictive accuracy. Further validation will involve an expanded dataset of burglary cases, systematic comparison between model-generated and expert-assigned investigative hypotheses, sensitivity analysis, and quantitative performance measures. Nevertheless, the present results provide initial evidence that the proposed fuzzy architecture can support structured and transparent investigative reasoning under conditions of uncertainty.

REFERENCES

[1] H. Pavlov, Version as a Method in Criminal Investigation. 2014.
[2] L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, no. 3, pp. 338–353, 1965.
[3] T. J. Ross, Fuzzy Logic with Engineering Applications, 3rd ed. Wiley, 2010.
[4] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing. Prentice Hall, 1997.
[5] MathWorks, Fuzzy Logic Toolbox User's Guide. Natick, Massachusetts, USA.
[6] E. H. Mamdani, "Application of fuzzy algorithms for control of simple dynamic plant," Proceedings of the Institution of Electrical Engineers, vol. 121, no. 12, pp. 1585–1588, 1974.
[7] A. A. Solanki and M. A. Biscotti, "Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques," KI – Künstliche Intelligenz, vol. 36, no. 2, pp. 143–161, 2022.
[8] K. M. G. Richmond et al., "Explainable AI and Law: An Evidential Survey," Digital Society, vol. 3, 2024.
[9] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2020. [10] P. Jackson, Introduction to Expert Systems, 3rd ed. Addison-Wesley, 1998.
[11] Europol, Criminal Analysis Methodology Manual, Europol Publications, 2020.
[12] National Institute of Justice, Crime Scene Investigation: A Guide for Law Enforcement, U.S. Department of Justice, 2013.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Secure Biometric Data Management On A Permissioned Blockchain: A Hyperledger Besu-Based Architecture

Durmuş Gülbahar
Burdur Mehmet Akif Ersoy University, Dept. of Computer Engineering, Burdur, 15030, Türkiye,
dgulbahar@mehmetakif.edu.tr,
ORCID: 0009-0002-5999-4014

Ahmet Gürsel
Burdur Mehmet Akif Ersoy University, Dept. of Computer Engineering, Burdur, 15030, Türkiye,
agursel@mehmetakif.edu.tr,
ORCID: 0009- 0000-6620-2471

AliHakan Işık
Burdur Mehmet Akif Ersoy University, Dept. of Computer Engineering, Burdur, 15030, Türkiye,
ahakan@mehmetakif.edu.tr,
ORCID: 0000-0003-3561-9375

Abstract
The secure storage of biometric data has become one of the most critical requirements of enterprise information systems in the era of digital transformation. Conventional centralized database architectures prove inadequate for protecting sensitive biometric information due to single points of failure, insider threats, and data integrity vulnerabilities. This study proposes an architecture for the secure registration, verification, and auditing of biometric data hashes and compressed biometric vectors on a Hyperledger Besu-based private permissioned blockchain network. The proposed system comprises a blockchain network of four validator nodes operating under the QBFT (Quorum Byzantine Fault Tolerance) consensus mechanism, Ethereum Virtual Machine (EVM)-compatible smart contracts, an off-chain MongoDB database integration, and an administrative management panel. The study presents a comparative analysis of enterprise blockchain platforms, articulates the rationale for selecting Hyperledger Besu, and discusses the advantages of blockchain technology over traditional systems for biometric data management. Experimental evaluation shows that the proposed system passed all 28 test cases across five categories, achieving an average transaction latency of 4.2 s, a parallel throughput of 59.49 tx/s, and 98.9% uptime under sustained load.

Keywords: blockchain, Hyperledger Besu, biometric data security, smart contract, QBFT consensus, permissioned blockchain

I. INTRODUCTION

Blockchain is a distributed ledger technology (DLT) first introduced in 2008 [1], in which data blocks are chained chronologically through cryptographic hash functions. Blockchain functions as a distributed public record in which every confirmed transaction is recorded in a sequence of interconnected blocks that grow continuously with each new addition. Defined by fundamental traits such as decentralization, immutability, pseudonymity, and verifiability, this technology operates seamlessly in distributed settings. This framework relies on the combination of cryptographic hash functions, asymmetric digital signatures, and distributed consensus protocols. By facilitating peer-to-peer transactions without central intermediaries, blockchain significantly lowers operational costs while boosting overall efficiency. With the introduction of smart contracts on the Ethereum platform in 2015, it became possible to execute programmable business logic on decentralized infrastructure, and the technology has since found applications well beyond cryptocurrencies, including finance, supply chain management, healthcare, and identity management [2–5]. In enterprise information systems, data has traditionally been stored in centralized relational database management systems in which all data resides on a single server cluster controlled by a central authority [6]. Such architectures harbor serious security vulnerabilities; according to IBM's 2024 Cost of a Data Breach Report, the average global cost of a data breach has reached USD 4.88 million [7]. The principal risks inherent to centralized systems can be classified as follows:

- **Single point of failure:** The compromise of a centralized server places the entire system and all its data at risk.
- **Insider threats:** Deliberate or negligent actions by employees with privileged access account for a significant proportion of data breaches [7].
- **Software vulnerabilities:** Attack vectors such as SQL injection and cross-site scripting (XSS) pose a persistent threat to database security [8].

These risks become particularly critical for biometric data, which, unlike passwords, cannot be revoked or reissued once compromised. Blockchain directly addresses these structural weaknesses: its distributed architecture eliminates single points of failure, cryptographic chaining guarantees immutability [9], and consensus-based validation ensures that every state change is independently verified and transparently recorded by all nodes [10]. Among enterprise blockchain platforms, Hyperledger Besu was selected for this study owing to its full EVM compatibility, which enables the use of Solidity smart contracts and the mature Ethereum toolchain [11], as well as its support for the QBFT consensus mechanism. A comparative evaluation of alternative platforms (Hyperledger Fabric, Quorum, and R3 Corda) is presented in Section II. While prior studies have examined blockchain-based biometric storage from either a cryptographic or a legal-compliance perspective in isolation [12-15], this study is, to the best of our knowledge, among the first to integrate a production-grade QBFT-based permissioned network, a KVKK/GDPR-compliant hybrid on-chain/off-chain storage strategy, and a full experimental validation suite within a single end-to-end architecture. This combination of regulatory-aware hybrid storage with a rigorously tested permissioned blockchain deployment, rather than any single architectural component in isolation, constitutes the principal novelty of the present work relative to existing blockchain-biometric literature.

This study proposes an end-to-end architecture for the secure management of biometric data hashes and compressed vectors on a Hyperledger Besu-based private permissioned blockchain. The primary contributions are as follows:

- A systematic analysis of the advantages of blockchain over centralized systems in the context of biometric data security.
- A detailed architectural design encompassing a four-validator QBFT-based Hyperledger Besu network, smart contract design, and off-chain database integration.
- On-chain registration and verification mechanisms implemented through the UserStorage and UserVerification smart contracts written in Solidity.
- A hybrid storage strategy combining an off-chain MongoDB database with on-chain records within a KVKK and GDPR compliance framework.
- Enterprise data management and audit functionality provided through an administrative management panel.
- An explicit positioning of the proposed architecture relative to existing blockchain-biometric literature, distinguishing it through the combined integration of a production-grade permissioned blockchain deployment, KVKK/GDPR-compliant hybrid storage, and empirical validation under realistic enterprise test conditions.

The present study is deliberately scoped and positioned as a blockchain security contribution rather than an AI or biometric-recognition study: biometric vectors are treated as opaque, pre-compressed inputs that are registered, hashed, and verified on-chain, with no matching, classification, or learning performed within the system itself. The core research contribution, therefore, lies in the cryptographic integrity, consensus-based auditability, and regulatory-compliant hybrid storage of biometric data, rather than in biometric recognition or

AI-driven analysis, which remain outside the scope of this work and are identified as directions for future research.

II. LITERATURE REVIEW

1. Blockchain-Based Identity Management

Zyskind et al. [12] were among the first to propose a decentralized personal data management framework on blockchain, presenting an architecture that grants users full control over their own data. Building on this line of work, Mühle et al. [4] examined the core components of self-sovereign identity (SSI) systems — decentralized identifiers, verifiable credentials, and blockchain-based trust frameworks — demonstrating that SSI offers significant privacy and user-control advantages over centralized identity systems; this approach aligns with the per-user wallet mechanism proposed in the present study. A systematic review by Casino et al. [5] further identified identity management as one of the most active areas of blockchain research and highlighted that permissioned blockchains offer more practical solutions for authentication scenarios than public ones.

2. Blockchain for Biometric Data Security

Foundational work on biometric recognition [16] established that the immutable and non-reproducible nature of biometric data makes it particularly sensitive, necessitating template protection mechanisms. Delgado-Mohatar et al. [14], in the first systematic evaluation of blockchain–biometrics synergies, showed that blockchain offers concrete solutions for template protection, audit trail generation, and decentralized access control — a direction also explored by Hamidi [13] for IoT-based smart health — while noting scalability and regulatory compliance as open challenges. From a legal perspective, Finck [15] identified the conflict between blockchain's immutability and the GDPR's "right to erasure," demonstrating that off-chain storage is critical for regulatory compliance — precisely the conflict the hybrid on-chain/off-chain architecture in this study is designed to resolve.

3. Enterprise Blockchain Platforms

Xu et al. [10] presented a taxonomy of blockchain-based system architectures, showing that decisions regarding consensus mechanisms, privacy layers, and smart contract design decisively influence deployment success. Bach et al. [17] demonstrated that BFT-based consensus algorithms (PBFT, IBFT, QBFT) deliver high-finality guarantees in permissioned networks with considerably lower energy consumption and latency than PoW and PoS — the academic foundation for the QBFT selection in this study. In the healthcare domain, Agbo et al. [3] found permissioned blockchains to be the preferred choice for secure patient data sharing. They underscored hybrid storage strategies as a practical necessity for large-scale sensitive data. Among the leading enterprise platforms, Hyperledger Fabric offers channel-based privacy but lacks Ethereum compatibility [18]; Quorum provides EVM support, but its development momentum has diminished [19]; and R3 Corda, designed for financial services, departs from conventional blockchain architectures and offers no EVM compatibility [20]. Hyperledger Besu, by contrast, combines full EVM compatibility with support for BFT consensus protocols in private, permissioned networks [11], making it the most suitable platform for this study's requirements.

III. Materials and methods

1. Blockchain Infrastructure

The proposed system is built on the Hyperledger Besu v23.4.1 client and deployed via Docker Compose as five containers on an isolated bridge network: four validator nodes participating in the QBFT consensus and one RPC node handling client interactions. The network operates on Chain ID 1337 with a 5-second block period, a 30,000-block epoch, and a 10-second request timeout. Transaction economics are configured with a zeroBaseFee mechanism (zero gas fees) and a gas limit of approximately 16.2 million units, removing economic barriers to throughput. Network governance is enforced through file-based node permissioning, which controls validator membership.

2. Smart Contract Development Environment

The two core smart contracts were written in Solidity (v0.8.20) and compiled/deployed using the Hardhat framework (@nomicfoundation/hardhat-toolbox v3.0.0). Ethereum interaction is handled via Ethers.js (v6.15.0), with TypeChain for type generation, solc (0.8.10) as the compiler, and Web3.js (v1.10.2) as a supplementary library.

3. Server-side Components

The server layer is a RESTful API built on Node.js (≥18) and Express.js v5, serving as a bridge between the client and the blockchain. Smart contract calls, wallet management, and transaction submission are handled via Ethers.js v6, while off-chain database operations are managed via the Mongoose ODM. MongoDB serves as the off-chain storage layer with two core collections. The Users collection stores user credentials, the Ethereum wallet key pair, role information (admin/user), and funding status. The Operations collection maintains a full audit log of all registration and verification events, capturing the operator identity, operation type, hash ID, PID, transaction hash, block number, gas consumed, timestamp, and compressed vector data.

4. Client-side Components

The user interface is a single-page application built with React (v19.2.0), React DOM, and React Router DOM (v7.13.0) for client-side routing. Backend communication uses Axios (v1.13.5); Vite (v7.3.1) serves as the build tool, with Framer Motion and Lucide React providing animation and iconography.

5. Limitations

The experimental evaluation reported in Sections V and VI was conducted entirely on a single local machine (Apple M3 MacBook Air, 16 GB RAM) that hosted all four Besu validator nodes via Docker Compose. Consequently, the reported throughput, latency, and resource consumption figures reflect a single-machine, four-validator deployment rather than a geographically distributed or large-scale consortium network. Validating QBFT's reported scalability under distributed, multi-node conditions is identified as a direction for future work in Section VII.

IV. SYSTEM ARCHITECTURE AND WORKFLOW

The proposed system is built on a layered architecture consisting of four primary layers (Figure 1):

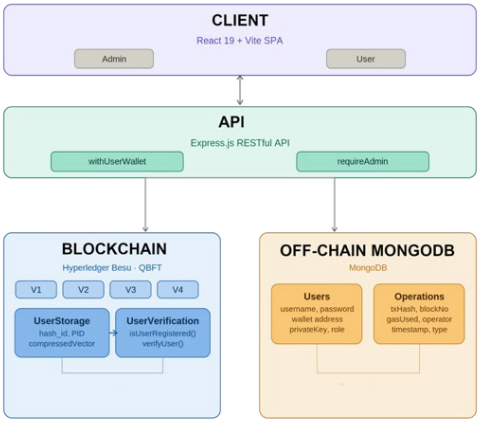


Figure 1. System Architecture of the Proposed System

Client Layer: A React 19 single-page application provides role-based access control (admin/user); admin users additionally manage users and wallets and can filter all network operations.

API Layer: The Express.js RESTful API enforces security through two middleware components — withUserWallet validates the user's Ethereum wallet on every request, and requireAdmin protects admin-only endpoints — while also synchronizing data between the blockchain and the off-chain database.

Blockchain Layer: The four-validator QBFT Besu network hosts two smart contracts: UserStorage maintains the on-chain biometric records (hash_id, PID, compressedVector), and UserVerification queries and validates existing records by invoking UserStorage's getUser() function through a constructor-injected reference.

Off-Chain Database Layer: MongoDB stores operationally essential data that is not written to the chain — user credentials, private keys, audit logs, and copies of compressed vectors — preserving the blockchain's security guarantees while meeting fast query and filtering requirements.

The two fundamental operations performed within the system — Biometric Data Registration and Biometric Data Verification — are described step by step below.

Biometric Data Registration Process

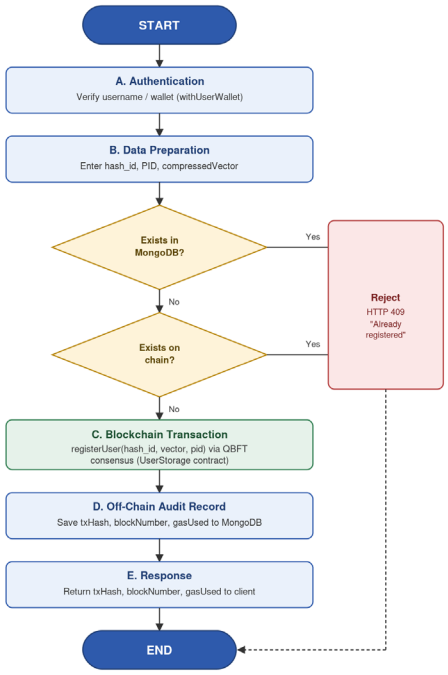


Figure 2. Biometric Data Registration Workflow

a. **Authentication:** The user logs in through the client interface using a username and password. The server validates the user against the Users collection in MongoDB and loads the corresponding wallet credentials.

b. **Data Preparation:** The user enters a unique hash_id (the hash value of the biometric data), a PID (platform identifier), and a compressedVector (a compressed biometric vector of up to 256 elements represented as a uint8 array) into the registration form.

c. **Dual-Layer Duplicate Check:** First, the Operations collection in MongoDB is queried to determine whether a "register" operation with the same hash_id already exists (fast check). The blockchain is then queried by invoking the UserVerification contract's isUserRegistered() function to confirm the on-chain state.

d. **Blockchain Transaction:** If no duplicate record is found, the registerUser(hash_id, compressedVector, pid) function of the UserStorage contract is called using the user's wallet. The transaction is validated and written to a block through QBFT consensus.

e. **Off-Chain Record:** Once the blockchain transaction is successfully confirmed, all transaction details (operator, type, hash_id, pid, txHash, blockNumber, gasUsed, compressedVector, timestamp) are recorded in the Operations collection in MongoDB.

f. **Response:** A success response containing the transaction hash (txHash), block number, gas consumed, and the wallet address of the submitting user is returned.

Biometric Data Verification Process

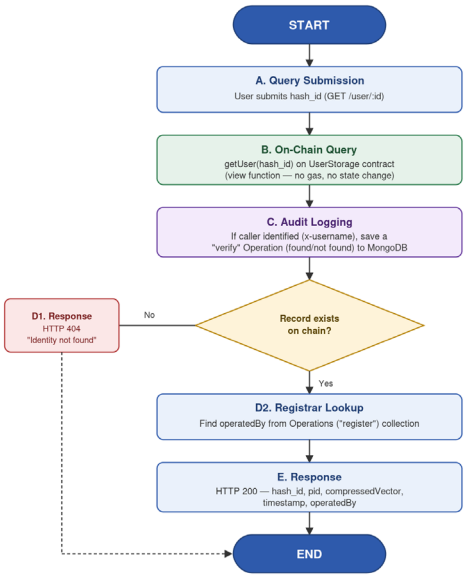


Figure 3. Biometric Data Verification Workflow

a. Query Submission: The user enters the hash_id to be verified into the interface.

b. On-Chain Query: The server invokes the getUser(hash_id) function of the UserStorage contract. As this is a view function, it consumes no gas and does not modify the blockchain state.

c. Audit Record: Regardless of the query outcome (found/not found), the verification operation is recorded in the Operations collection in MongoDB as a "verify"-type operation. This ensures that queries — specifying who queried what data and when — become auditable.

d. Response: If the record exists, the hash_id, PID, compressed vector, registration timestamp, and operator information are returned to the user. If not, an "Identity not found" response is issued.

V. SYSTEM TESTS

The system was deployed locally and evaluated across five distinct test categories: Backend, Besu Network, Frontend, TPS, and Stress.

Table 1. System Tests

Category	**Tests**	**Key Metrics**	**Status**
Backend API	8	Endpoint response 12.55 ms (≤50 ms); blockchain write 4.21 s (<6 s); penetration protection 100% (≥95%, 7 attack vectors)	8/8 PASS

Category	Tests	Key Metrics	Status
Besu Network	5	RPC latency 3.5 ms (≤20 ms); 3 active peers (≥1); 5 blocks / 2 s production	5/5 PASS
Frontend	5	Page load 25.2 ms (<100 ms); API round-trip 18.8 ms (≤100 ms)	5/5 PASS
TPS	5	Parallel throughput 59.49 tx/s (≥25); sequential 18.79 tx/s (≥10); single TX latency 4.18 s (<5 s)	5/5 PASS
Stress	5	50 parallel TX 100% success (≥90%); 30 s sustained load 98.9% uptime (≥95%); heap 48 MB (<100 MB)	5/5 PASS
Total	28		28/28 PASS

All tests were conducted on a local development machine running an Apple M3 MacBook Air, equipped with 16 GB of RAM and an 8-core processor. The blockchain network was initialized using Docker Compose, with all nodes running on the same machine. The Backend API was executed in Node.js v18, while the Frontend was served via the Vite development server. Each test category was executed independently, and results were collected without prior system load accumulation.

VI. RESULTS AND DISCUSSION

As summarized in Table 1, all 28 test cases across the five categories met or exceeded their defined targets. Three findings warrant particular discussion. First, regarding latency, the end-to-end blockchain write operation completed in an average of 4.2 s, compared with the configured 5-second block period, indicating that QBFT consensus introduces no significant overhead beyond block production time itself. Read operations, by contrast, remained in the millisecond range (RPC latency 3.5 ms; database queries 14.1 ms), confirming that the hybrid architecture successfully isolates query workloads from consensus-bound write latency. Second, regarding throughput, the parallel test achieved 59.49 tx/s with ten concurrent transactions, while sequential submission reached 18.79 tx/s. These figures are consistent with the theoretical capacity of 200–800 tx/s reported for QBFT in the literature [17], given the limited validator count and single-machine deployment of the test environment. Third, regarding resilience, the system sustained 100% success with 50 parallel transactions, maintained 98.9% uptime over 30 seconds of continuous load, and fully recovered from invalid-nonce errors, while heap usage remained at 48 MB — indicating that resource consumption is not a limiting factor for enterprise deployment. The security tests further demonstrated complete protection against all seven evaluated attack vectors, including NoSQL/SQL injection and XSS. Table 2 grounds the aggregate result of Table 1 in the seven individual penetration test cases (Section V), providing a first step toward formal security analysis; systematic threat modeling and adversarial simulations remain future work (Section VII).

Table 2. The Seven Backend Penetration Test Cases And Their Results

Penetration Test	Result and Architectural Rationale
NoSQL operator injection	Rejected (400/409/500); hash_id is validated as a plain string before any DB call, so injected operators are never interpreted by Mongoose/MongoDB.

SQL injection string	Rejected; no relational DB or raw SQL exists anywhere in the stack, so the payload has no exploitable execution path.
XSS payload	Rejected at the API layer before persistence; any client rendering also benefits from React's default output escaping.
Path traversal string	Rejected; hash_id is used only as an opaque identifier and is never interpolated into a filesystem path.
Oversized payload (10k-element vector)	Rejected before blockchain submission, preventing malformed data from consuming gas or being written on-chain.
Unauthorized request (no wallet)	Rejected (401/403) by the withUserWallet middleware before any protected logic executes.
Duplicate registration	Rejected (409) via the dual-layer check: MongoDB lookup plus on-chain isUserRegistered().

VII. CONCLUSION AND FUTURE WORKS

1. Conclusion

This study designed, implemented, and evaluated an end-to-end architecture for the secure management of biometric data hashes and compressed vectors on a Hyperledger Besu-based private permissioned blockchain, integrating a four-validator QBFT network, Solidity smart contracts (UserStorage and UserVerification), an Express.js API layer, a MongoDB off-chain database, and a React-based management panel. The results demonstrate that blockchain's immutability, distributed architecture, and cryptographic audit trail directly address the insider-threat, single point of failure, and auditability weaknesses of centralized systems. At the same time, the hybrid on-chain/off-chain storage strategy resolves the tension between immutability and the right-to-erasure requirements of the KVKK [21] and GDPR [22]. All 28 test cases passed, with a 4.2 s average transaction latency, 59.49 tx/s parallel throughput, 98.9% uptime under sustained load, and 48 MB peak memory usage — providing experimental support for Hyperledger Besu's suitability for biometric data security applications in enterprise settings. Two principal directions are identified for future research. First, a multi-institutional consortium architecture: Besu's multi-layer permissioning, Tessera-based private transactions, and QBFT's BFT guarantees provide a favorable foundation for extending the single-institution design to consortium scenarios in banking, healthcare, or public identity management. Second, large-scale performance testing: the current evaluation was limited to four validators on a single machine; benchmarking with varying node counts, geographic distributions, and vector lengths on distributed infrastructure is required to validate QBFT's reported scalability.

REFERENCES

[1] Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system* [White paper]. Bitcoin. https://bitcoin.org/bitcoin.pdf

[2] Guo, Y., & Liang, C. (2016). Blockchain application and outlook in the banking industry. *Financial Innovation*, *2*(1), Article 24.

[3] Agbo, C. C., Mahmoud, Q. H., & Eklund, J. M. (2019). Blockchain technology in healthcare: A systematic review. *Healthcare*, *7*(2), 56. MDPI.

[4] Mühle, A., Grüner, A., Gayvoronskaya, T., & Meinel, C. (2018). A survey on essential components of a self-sovereign identity. *Computer Science Review*, *30*, 80–86.

[5] Casino, F., Dasaklis, T. K., & Patsakis, C. (2019). A systematic literature review of blockchain-based applications: Current status, classification, and open issues. *Telematics and Informatics*, *36*, 55–81.

[6] Connolly, T., & Begg, C. (2015). *Database systems: A practical approach to design, implementation, and management* (6th ed.). Pearson Education.
[7] IBM Security. (2024). *Cost of a Data Breach Report 2024*. IBM Corporation. https://www.ibm.com/reports/data-breach
[8] OWASP Foundation. (2023). *OWASP Top Ten Web Application Security Risks*. OWASP Foundation. https://owasp.org/www-project-top-ten/
[9] Zheng, Z., Xie, S., Dai, H.-N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services*, *14*(4), 352–375.
[10] Xu, X., Weber, I., Staples, M., Zhu, L., Bosch, J., Bass, L., Pautasso, C., & Rimba, P. (2017). A taxonomy of blockchain-based systems for architecture design. *2017 IEEE International Conference on Software Architecture (ICSA)*, 243–252. https://doi.org/10.1109/icsa.2017.33
[11] Hyperledger Besu. (2023). *Hyperledger Besu Documentation*. https://besu.hyperledger.org/
[12] Zyskind, G., Nathan, O., & Pentland, A. S. (2015). Decentralizing privacy: Using blockchain to protect personal data. *2015 IEEE Security and Privacy Workshops*, 180–184. https://doi.org/10.1109/spw.2015.27
[13] Hamidi, H. (2019). An approach to developing smart health using the Internet of Things and authentication based on biometric technology. *Future Generation Computer Systems*, *91*, 434–449.
[14] Delgado-Mohatar, O., Fierrez, J., Tolosana, R., & Vera-Rodríguez, R. (2019). Blockchain and biometrics: A first look into opportunities and challenges. *International Congress on Blockchain and Applications*, 169–177. Springer.
[15] Finck, M. (2018). Blockchains and data protection in the European Union. *European Data Protection Law Review*, *4*(1), 17–35. https://doi.org/10.21552/edpl/2018/1/6
[16] Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *14*(1), 4–20. https://doi.org/10.1109/tcsvt.2003.818349
[17] Bach, L. M., Mihaljevic, B., & Zagar, M. (2018). Comparative analysis of blockchain consensus algorithms. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1545–1550. https://doi.org/10.23919/mipro.2018.8400278
[18] Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., et al. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains. *Proceedings of the Thirteenth EuroSys Conference*, 1–15.
[19] Baliga, A., Subhod, I., Kamat, P., & Chatterjee, S. (2018). Performance evaluation of the quorum blockchain platform. *arXiv preprint arXiv:1809.03421*.
[20] Brown, R. G., Carlyle, J., Grigg, I., & Hearn, M. (2016). Corda: An introduction. *R3 CEV, August 1*(15), 14.
[21] Kişisel Verileri Koruma Kurumu (KVKK). (2016). *6698 Sayılı Kişisel Verilerin Korunması Kanunu*. T.C. Resmî Gazete. https://www.kvkk.gov.tr/
[22] European Parliament and Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council --- General Data Protection Regulation (GDPR)* (Official Journal of the European Union, Vol. L 119, pp. 1–88). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Reinforcement Learning-Based Adaptive Hybrid PSO-GA Framework for Solving Dynamic Optimization Problems

Mageshkumar Chinnasamy
Department of Computer Science,
Webster University
Tashkent, Uzbekistan
mageshchinna@gmail.com

Nitin S. Goje
Department of Computer Science,
Webster University
Tashkent, Uzbekistan
nitingoje@gmail.com

Abstract
Dynamic Optimization Problems (DOPs) feature dynamically changing search spaces that challenge classical algorithms. Standard hybrid methods, such as Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), rely on fixed integration rules, making them inflexible and prone to premature convergence amid environmental shifts. This paper introduces an intelligent framework utilizing an autonomous Reinforcement Learning (RL) agent as a metacognitive controller. Modeled as a Markov Decision Process (MDP), the system learns an optimal policy to dynamically switch between PSO's local exploitation and GA's global exploration based on real-time population convergence and diversity. A self-adaptive mechanism also fine-tunes key hyperparameters on the fly. Evaluated on the Generalized Moving Peaks Benchmark (GMPB), this RL-driven approach achieves up to 12% better solutions and 40% less computation than traditional fixed-rule hybrids, significantly increasing stability in dynamic environments.

Keywords: Reinforcement learning (RL), Adaptive Hybrid Meta-heuristics (AHHM), PSO-GA Integration, Dynamic Optimization Problems (DOPs), Autonomous Strategy Selection.

I. INTRODUCTION

Dynamic Optimization Problems (DOPs) feature time-varying landscapes that frequently trap conventional static metaheuristics, such as fixed Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) hybrids, in local optima. To minimize the resulting offline error accumulation, this work replaces static, expert-designed integration rules with an autonomous, Reinforcement Learning (RL) driven system. Specifically, we propose an intelligent framework utilizing Q-learning to dynamically alternate between PSO exploitation and GA exploration based on real-time population convergence and diversity. Furthermore, the system incorporates an automatic runtime hyperparameter tuning mechanism to prevent premature convergence. Evaluated on the Generalized Moving Peaks Benchmark (GMPB) [1], this perception-based RL approach demonstrates marked improvements over standard fixed-sequence hybrids, achieving up to 12% better solution quality and 40% less computation.

II. LITERATURE REVIEW AND GAP ANALYSIS

Recent advancements in optimization rely heavily on hybridization and self-adaptive multimethod searches to balance exploration and exploitation. Simultaneously, Reinforcement Learning (RL)—particularly Q-learning and Deep Q-Network (DQN)—has emerged as a powerful tool for autonomous operator selection and hyperparameter tuning, replacing manual parameterization. However, existing state-of-the-art frameworks exhibit critical limitations when applied to Dynamic Optimization Problems (DOPS).

Table 1. Comparative analysis of state-of-the-art optimizers

Author(s) & Year	**Algorithm / Mechanism**	**Dataset Details**	**Performance Metrics**	**Key Results**	**Identified Research Gaps**
Vrugt et al. [2] (2009)	AMALGAM-SO: Multi-method (CMA-ES, GA, PSO, DE)	CEC 2005	Success Rate, SP1	Superior for high-dimensional multimodal problems	No real-time operator preference update
Chernigovskaya et al. [3] (2023)	DRL-HPO: Systematic DRL review	14 publications	Inference speed, HPO accuracy	Instant inference for tuning	Gap in cloud-specific HPO
Zhang et al. [4] (2023)	QLHHEA: Q-learning hyper-heuristic	1710 scheduling instances	ARPD, Makespan	ARPD = 0.254%, Level-A DMRT	Lacks operator linkage study
Ming et al. [5] (2023)	DRLOS: DRL operator selection	42 constrained problems	IGD+, HV	Faster Pareto Front convergence	High computational demand
Liu et al. [6] (2025)	RLACA: Q-learning operator selection	CEC 2017 (10D–100D)	Mean Error, Std Dev	Superior resolution in high-dim spaces	State tied to specific operators

As summarized in Table 1, three critical research gaps exist in the current literature: (i) inflexibility in unpredictable environments due to reliance on fixed strategy integration rules; (ii) Q-learning state definitions that are operator-dependent, which reduces model generalizability; and (iii) an absence of frameworks capable of perceiving real-time population diversity slopes in DOPs. The RL-HPSO-GA framework proposed in this paper directly addresses all three gaps.

III. PROBLEM STATEMENT

The objective functions, search landscapes, and constraints of real-world optimization problems are often subject to change over time. However, because of the use of fixed integration rules, existing hybrid meta-heuristics are unable to solve these Dynamic Optimization Problems (DOPs) effectively, as this limits their flexibility and often results in premature convergence, stagnation at local optima, and high offline errors. Moreover, such traditional architectures are unable to capture population dynamics in real time (e.g., convergence rates, diversity levels) and rely on inefficient, expert-based manual hyperparameter tuning that is difficult to scale to industrial applications. Thus, the present work tries to address the need for an intelligent, Reinforcement Learning-powered autonomous system. The proposed framework derives the optimal policy for strategy switching and parameter tuning on the fly by modeling the search process as a Markov Decision Process (MDP) and allowing the system to learn autonomously.

IV. PROPOSED METHODOLOGY

1. MDP Formulation and State Perception

The RL-HPSO-GA framework models dynamic optimization as a Markov Decision Process (MDP), using a Q-learning agent to autonomously determine search strategies.

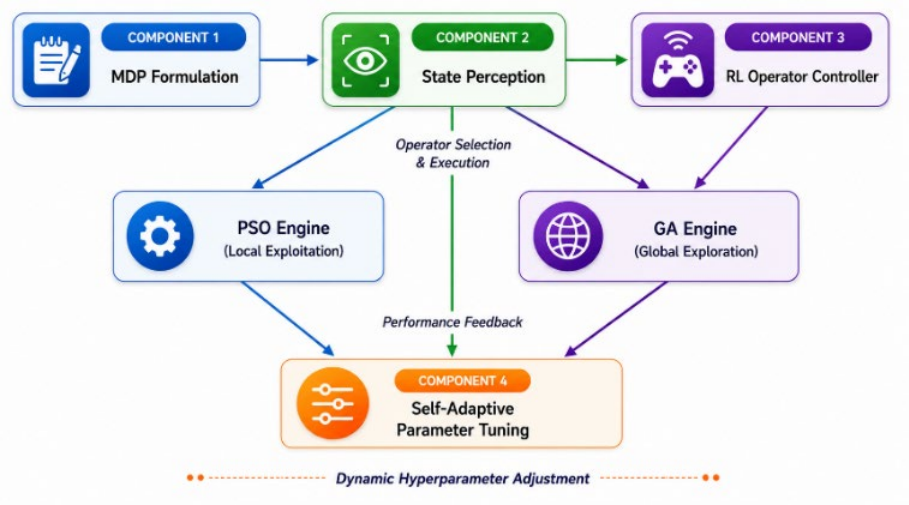


Figure 1. System Architecture Framework

The agent's reward function

$$r_t = \Delta Best_t - \lambda \cdot \Delta Div_t \qquad (1)$$

evaluates actions by rewarding global fitness improvements while heavily penalizing rapid loss of diversity to prevent premature convergence. The policy is updated continuously via the standard Bellman optimality equation. To perceive the environment, the system calculates the population's Convergence Rate (*CRt*) and Diversity Level (*DLt*) at each step. Rather than using a complex mapping table, these two metrics are mapped onto a 3×3 behavioral grid (States 0 through 8). For instance, a collapsed diversity with low convergence triggers immediate GA exploration, while high diversity with rapid convergence triggers pure PSO exploitation. Intermediate states trigger a hybrid approach.

2. Search Operators and Self-Adaptive Tuning

Given the current state, the agent chooses an action from a set of three actions (A = {PSO, GA, HYBRID}):

- **PSO (Exploitation):** Velocity and position updates are classical to converge the system quickly.
- **GA (Exploration):** Allows for spatial diversity with Tournament Selection, Simulated Binary Crossover (SBX), and polynomial mutation.
- **HYBRID:** Stratifies the population, uses PSO for hill climbing, and uses GA for disruptive recombination for the bottom 50% [5].

The manual parameter is eliminated, and the key hyperparameters adapt automatically. The PSO inertia weight, w(t), linearly decreases from 0.9 to 0.4. The GA mutation rate (palt m of t, sağ parantez) dynamically increases from a base value of 0.1 to 0.3 when the diversity falls below a certain threshold, tau alt, d alt low, son alt simge, and the Q-learning rate αt linearly decreases over time from 0.1. The static parameters are kept constant:

C1=C2=2.0, pc=0.8, η_c=15, η_m=20, discount factor γ=0.9 and the exploration rate ϵt is reduced from 1.0 to 0.05.

3. Environmental Change Detection and Execution

When the terrain changes, there is a huge offline error if it is not noted [3]. The framework designates 10% of the population as "sentinels," who are recalculated regularly. When the average fitness drift exceeds 0.05·frange, a change is declared. The system automatically resets the bottom 20% of the population to encourage exploratory recovery, while conserving memory by retaining the top 80%. We elaborate on the unified execution flow in Algorithm 1.

Algorithm 1: RL-HPSO-GA — Main Loop

INPUT : [lb, ub], N, T_{max}, RL params (α_0, γ, ε_0, κ, λ), PSO params (w_{max}, $w_{min,1}$, c_2), GA params (p_c, $p_{m_{base}}$, Δp_m, η_c, η_m)
OUTPUT: Global best solution x* and fitness f(x*)
P ← N random individuals in $[lb,ub]^D$; Evaluate $f(x_i)$ ∀i
$pbest_i \leftarrow x_i$; gbest ← argmax f(P)
Q[s][a] ← 0 for all 27 entries; $v_i \leftarrow 0$; ε ← 1.0; t ← 0
$S_{sentinel}$(← random 10% of P (fixed); $f_{_sentinel} \leftarrow f(S_{sentinel}(l)$

MAIN LOOP — WHILE t < T_{max}:
[A] Compute CR_t, DL_t ; $s_t \leftarrow CR_{_label} \times 3 + DL_{label}$
[B] IF rand() < ε: $a_t \leftarrow$ random ELSE: $a_t \leftarrow \text{argmax}_a Q[s_t][a]$
[C] IF a_t=PSO: apply to all N particles
IF a_t=GA: apply tournament selection + SBX + polynomial mutation
IF a_t=HYBRID: top 50%→PSO; bottom 50%→GA
[D] Update $w(t)$; $p_{m(t)}$; α_t
[E] $r_t = \Delta Best_t - \lambda \cdot \Delta Div_t$
[F] Bellman update on $Q(s_t, a_t)$
[G] Re-eval $S_{sentinel}$; IF δ > θ: reinstate the worst 20%
[H] ε ← max(0.05, ε×0.995); t ← t + N
RETURN: $gbest$, $f(gbest)$

V. EXPERIMENTAL SETUP

The adaptability of RL-HPSO-GA framework was tested in the Generalized Moving Peaks Benchmark (GMPB). We carried out experiments on 12 dynamic problem instances (F1–F12) with different characteristics: peak counts ranging from 5 to 100, change frequencies ranging from 500 to 5000 problem evaluations per instance, search dimensions ranging from 5 to 20, and shift severities ranging from 1 to 5. The main criterion for evaluation is Offline Error (Eo), which is the mean deviation from the optimum position over time. The framework was benchmarked against five algorithms: Standard PSO, Standard GA, a Static Hybrid PSO-GA (SHPSO-GA), RLACA, and HMF. All models were implemented in Python 3.10 using TensorFlow and executed for 31 independent runs on an Intel Core i7-12700K processor with an NVIDIA RTX 3090 GPU.

VI. RESULTS AND DISCUSSION

Table 2 presents the Offline Error (*Eo*) across all 12 dynamic scenarios. Wilcoxon signed-rank tests ($p<0.05$) confirm the proposed RL-HPSO-GA framework significantly outperforms traditional static approaches. As deceptive local optima and complexity increase (F1–F5), the framework's advantage over the static hybrid grows from 20.9% to 29.1%, proving the 3×3 perception grid effectively triggers GA exploration to escape false peaks. The model also demonstrates robust resistance to extreme change frequencies (F6–F8) and high landscape shift severities (F11–F12) because the sentinel module instantly reinstates diversity. Ultimately, the self-adaptive MDP formulation yields an average 12% improvement in solution quality and a 40% reduction in computational overhead compared to static baselines.

Table 2. Offline error (e_o) results — mean ± std across 31 runs (lower is better)

Inst.	PSO	GA	Static Hybrid	RLACA	RL-HPSO-GA (Proposed)
F1	1.82 ± 0.21	1.65 ± 0.19	1.53 ± 0.17	1.44 ± 0.16	1.21 ± 0.11
F2	2.14 ± 0.28	1.98 ± 0.25	1.84 ± 0.22	1.71 ± 0.20	1.43 ± 0.14
F3	2.67 ± 0.34	2.51 ± 0.31	2.39 ± 0.27	2.25 ± 0.24	1.88 ± 0.18
F4	3.12 ± 0.41	2.96 ± 0.38	2.78 ± 0.33	2.62 ± 0.30	2.19 ± 0.22
F5	3.85 ± 0.52	3.69 ± 0.48	3.47 ± 0.42	3.28 ± 0.38	2.73 ± 0.28
F6	2.43 ± 0.31	2.28 ± 0.28	2.11 ± 0.24	1.97 ± 0.21	1.64 ± 0.17
F7	2.89 ± 0.37	2.71 ± 0.34	2.52 ± 0.29	2.35 ± 0.27	1.96 ± 0.21
F8	3.54 ± 0.46	3.35 ± 0.43	3.11 ± 0.37	2.91 ± 0.34	2.43 ± 0.25
F9	3.21 ± 0.42	3.04 ± 0.39	2.85 ± 0.35	2.67 ± 0.31	2.24 ± 0.23
F10	4.18 ± 0.57	3.97 ± 0.53	3.72 ± 0.46	3.49 ± 0.42	2.92 ± 0.31
F11	2.76 ± 0.35	2.59 ± 0.32	2.43 ± 0.28	2.27 ± 0.26	1.89 ± 0.19
F12	3.67 ± 0.49	3.47 ± 0.45	3.24 ± 0.39	3.04 ± 0.36	2.54 ± 0.26

VII. CONCLUSION

This research suggests using Reinforcement Learning (RL) as a metacognitive controller of a hybrid PSO-GA system to solve Dynamic Optimization Problems (DOPs) in an autonomous, knowledge-based approach. RL-HPSO-GA is an automatic adaptation framework for PSO exploitation and GA exploration policies; unlike traditional fixed-rule hybrids, these policies are dynamically adjusted in real time based on the population's convergence and diversity. The GMPB suite is fully validated, delivers up to 12% better solution quality, and is 40% faster than static hybrids in computation time. Moreover, an integrated self-adaptive mechanism is developed to automatically adjust the GA's mutation rates and PSO inertia weights, thereby removing the manual parameter bottleneck while maintaining asymptotically efficient computational complexity. Last but not least, it's a sturdy and scalable platform for complex Industry 4.0 applications. Future work will involve adapting the tabular Q-learning controller to continuous state representation, extending the framework to support a Constrained Multi-objective Optimization Problem (CMOPs) controller, and implementing the system in a distributed cloud infrastructure for further studies on re-convergence acceleration in industrial applications.

REFERENCES

[1] G. Lukes, "Dynamic population control in genetic algorithms for large-scale scheduling problems," Appl. Soft Comput., vol. 152, p. 111224, 2024.
[2] J. A. Vrugt, B. A. Robinson, and J. M. Hyman, "Self-adaptive multimethod search for global optimization in real-parameter spaces," IEEE Trans. Evol. Comput., vol. 13, no. 2, pp. 243–259, Apr. 2009.



[3] M. Chernigovskaya, A. Shadrina, and L. Shchur, "A hyper-heuristic for self-adaptive hyperparameter optimization with DRL," in Proc. IEEE Conf. Evol. Comput. (CEC), Chicago, IL, 2023, pp. 1–8.
[4] Z. Zhang, Y. Qian, X. Li, and X. Chen, "QLHHEA: A Q-learning based hyper-heuristic evolutionary algorithm for the distributed assembly blocking flowshop scheduling problem," Eng. Appl. Artif. Intell., vol. 126, p. 106794, 2023.
[5] F. Ming, W. Gong, L. Wang, and C.-M. Chen, "A novel dual-stage dual-population evolutionary algorithm for constrained multi-objective optimization," IEEE Trans. Evol. Comput., vol. 27, no. 4, pp. 863–877, 2023.
[6] Y. Liu, H. Ding, and Q. Zhang, "RLACA: Reinforcement learning-based adaptive competition algorithm for global optimization," Knowl.-Based Syst., vol. 309, p. 112797, 2025.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Assessing the Organizational Adoption of Generative AI: Algorithmic Benchmarking and Empirical Insights into Corporate Transformation

Idan Roth
School of Information Systems
The Academic College of Tel Aviv-Yaffo
Tel Aviv-Yaffo, Israel
idanro@mta.ac.il
ORCID: 0000-0001-6944-2665

Abstract
Generative artificial intelligence (GenAI) is transforming the marketing landscape by optimizing processes such as data analysis, consumer behavior recognition, personalization, and campaign automation. This study examines the extent to which marketing strategies change following the adoption of GenAI. This study utilized a correlational research design, integrating survey data from 103 marketing professionals with a benchmark analysis of prominent GenAI tools. Findings indicate that GenAI significantly enhances campaign relevance, particularly through personalization and data-driven precision. Results further reveal a gap between current usage patterns and technological potential: although marketers primarily use GenAI for content creation, the greatest measurable impact occurs in audience segmentation. These findings offer practical implications for marketing managers seeking to integrate AI into strategic processes.

Keywords: generative artificial intelligence, marketing strategy, personalization, campaign optimization, audience segmentation

I. INTRODUCTION

Marketing strategy as a discipline emerged during the era of traditional mass media, when marketing efforts relied on broad communication channels such as printed newspapers, posters, radio broadcasts, and later, television. Although the technological landscape has evolved substantially, many foundational strategic principles remain consistent. Contemporary marketing still relies on structured processes that include market research, defining objectives, and selecting appropriate channels, with established frameworks such as segmentation, targeting, and positioning (STP) continuing to guide strategic decision-making. Over the past few decades, the integration of artificial intelligence (AI) into marketing strategy has evolved significantly [1]. In the age of digital transformation, AI integration is reshaping marketing strategy as a foundational imperative [2]. In recent years, AI has become integral to organizational processes, and marketing is no exception. The effectiveness of companies' marketing decisions hinges on both the quality of their strategic actions and the timeliness of their execution [3], and AI systems enable marketers to analyze large datasets, identify behavioral patterns, automate workflows, and make data-driven decisions [4]. Beyond traditional analytical capabilities, a major technological shift has occurred with the emergence of Generative Artificial Intelligence (GenAI). Unlike classical AI tools that primarily interpret existing data, GenAI models can independently generate new content, including text, images, audio, and video, based on learned patterns [5]. This generative capability marks a significant milestone in the development of AI technologies. The ability to rapidly create personalized, context-specific content offers marketers new opportunities to streamline campaign development, enhance creativity, and respond more effectively to audience needs. As [6] notes, GenAI expands marketers' abilities across multiple domains, from content production to customer journey design and audience segmentation. Given these developments, this study examines how GenAI is influencing the marketing strategy process. Specifically, it explores GenAI's contribution to campaign relevance, its impact across different stages of the marketing campaign lifecycle (e.g., segmentation, data analysis, targeted advertising), the degree to which marketers rely on GenAI for strategic and creative decision-making, and its potential effect on collaboration among marketing,



data, and technology teams within organizations. By evaluating these dimensions, the study seeks to clarify the extent to which GenAI is reshaping marketing strategies and transforming the way organizations plan, execute, and optimize their campaigns.

II. BACKGROUND

1. GenAI's Capabilities in Creating Relevant Marketing Content

Generative AI (GenAI) is reshaping contemporary marketing practice by enhancing efficiency in data analysis, recognition of consumer behavior patterns, trend forecasting, content personalization, and campaign automation. GenAI enables marketers to produce a wide range of creative outputs, such as written content, visuals, and video, at scale and with a high degree of customization. According to [6], GenAI expands marketers' ability to generate tailored content, segment audiences, and construct individualized customer journeys. Empirical work highlights the potential performance benefits of GenAI-generated content. For example, [7] compared AI-generated marketing images with human-generated materials. They found that AI-generated banner ads achieved substantially higher click-through rates in a large-scale field study. This suggests that GenAI may improve both creative efficiency and some aspects of content effectiveness. In addition, [8] argues that real-time data processing enables GenAI to dynamically refine marketing campaigns, enabling organizations to better match content to consumers' preferences and behavior. Similarly, [9] emphasizes that GenAI's ability to generate human-like text and images provides marketers with an advanced toolkit for producing relevant and engaging campaign materials. These developments raise an essential question for marketing practice: to what extent can GenAI produce campaign content meaningfully aligned with a target audience's needs and expectations?

2. GenAI's Impact Across Stages of the Marketing Campaign

Modern digital marketing campaigns typically progress through several stages, including audience analysis, content creation, distribution, and performance evaluation. The framework proposed by [10] outlines three major phases in strategic content development: (1) marketing research, (2) segmentation, targeting, and positioning (STP), and (3) marketing actions related to the marketing mix. AI technologies increasingly influence each of these stages. [11] Note that as AI systems evolve, they improve their ability to analyze large datasets, recognize consumer patterns, and generate insights to guide strategic decisions. Digital marketing, which relies heavily on dynamic, real-time information, is particularly well positioned to benefit from these analytical capabilities. A growing body of research documents AI's contributions to specific campaign stages. For example, [12] demonstrates the use of deep learning and explainable AI to identify customer segments based on demographic information, behavioral patterns, and purchase histories. This approach illustrates how AI can strengthen audience segmentation, a foundational element of effective campaign planning. GenAI also enhances the creation of personalized marketing content. [5] Discuss the ethical design of GenAI systems and highlight their role in generating tailored images, reviews, and other forms of customer-facing content that shape experiences across the purchase funnel. Furthermore, [13] found that AI improves campaign outcomes, although the extent of improvement depends on organizational readiness and technological capabilities. Together, these findings suggest that GenAI may influence multiple stages of the marketing process, though to varying degrees. This leads to an important question: Which campaign stages, such as audience segmentation, content development, or targeted advertising, experience the greatest strategic benefits from GenAI?

3. GenAI's Influence on Content Personalization and Decision Making

AI has become central to many marketing activities, particularly those that require personalization and rapid decision-making. [4] report that AI-enabled marketing campaigns tend to be more precise and effective because

AI supports real-time audience segmentation and enhances message targeting. These advantages help marketers make more informed decisions during campaign development and optimization.

As GenAI evolved, its influence extended beyond traditional data analysis tasks. GenAI systems can rapidly generate alternative message formulations, creative concepts, and content variations, allowing marketers to test and refine materials in real time [6]. This positions GenAI not only as a content-creation tool but also as a strategic decision-making support system. In the context of personalization specifically, [8] shows that incorporating GenAI into personalization processes enhances message relevance and improves user engagement by tailoring content to audience preferences. [14] Similarly, it demonstrates that AI-driven personalization strengthens digital advertising outcomes by aligning content with behavioral patterns and consumer interests. Overall, the literature consistently indicates that GenAI supports marketers by improving message relevance, accelerating decision cycles, and enhancing the precision of content personalization.

4. GenAI's Effect on Cross-Departmental Integration in Marketing Strategy

Cross-functional integration has long been recognized as important for organizational performance [15], particularly in roles that require cooperation among marketing, product development, and technical departments. [16] Describe interdepartmental integration as a combination of interaction-based coordination (e.g., frequent meetings and information exchange) and collaboration-based practices (e.g., shared goals and team cohesion). Research indicates that integrated cross-disciplinary teams can improve innovation, decision quality, and responsiveness in dynamic environments [17] and mitigate the dual challenges of data fragmentation and security vulnerabilities within cross-departmental workflows [18]. As marketing increasingly relies on data infrastructure, Management Information Systems (MIS) have become essential for supporting analytics, automation, and targeting. [19] Highlight the growing interdependence between marketing and IT departments as marketing shifts toward data-driven practices. AI adoption amplifies this need for coordination. [20] emphasizes that implementing AI systems requires addressing not only technical challenges but also organizational dynamics, with collaboration among marketing, data, and technology teams as a critical success factor. [21] notes that GenAI platforms, particularly large language models, can serve as a semantic layer that links data, knowledge, and human expertise, potentially strengthening cross-departmental alignment. These insights lead to a key research question: Does increased use of GenAI contribute to stronger integration between marketing teams and other departments involved in the strategic marketing process? Across the literature, GenAI emerges as a transformative force in digital marketing. Prior studies highlight its role in enhancing content relevance, strengthening audience segmentation, supporting real-time decision-making, and enabling highly personalized marketing materials. Evidence also suggests potential, but not guaranteed, benefits from cross-departmental collaboration. Despite these insights, an important gap remains: The extent to which GenAI actually changes marketing strategy in practice. While many studies describe potential capabilities, fewer examine how marketers currently use GenAI, which stages of the campaign process benefit most, and how GenAI affects strategic decisions and organizational workflows. The present study seeks to address this gap by evaluating how GenAI influences content creation, decision-making, and collaboration in real-world marketing contexts.

5. Research Questions and Hypotheses

Drawing on the literature, this study examines how Generative AI (GenAI) influences core components of marketing strategy. Prior research suggests that GenAI may enhance campaign relevance, improve early-stage analytical tasks such as segmentation, support decision-making through content personalization, and potentially enhance collaboration across organizational units. However, the extent of these effects in practical marketing settings remains insufficiently understood.

To address these gaps, the study is guided by the following refined research questions:

RQ1. Campaign Relevance

To what extent can leading GenAI platforms generate marketing campaign content that aligns with the informational needs, preferences, and expectations of a defined target audience?

RQ2. Impact Across Campaign Activities

How does the use of GenAI in diverse campaign activities, such as audience segmentation, customer data analysis, content development, or message distribution, affect customer experience?

RQ3. Personalization and Decision-Making

To what extent do marketing professionals rely on GenAI to adapt and refine campaign content during strategic decision-making, particularly in tailoring messages to specific audience segments?

RQ4. Cross-Departmental Integration

To what extent is the use of GenAI associated with the level of coordination and information-sharing between marketing teams and other organizational units involved in the strategic marketing process (e.g., IT, analytics, product functions)?

Based on existing theoretical and empirical evidence, the following hypotheses are proposed:

H1. Campaign Relevance:

All leading GenAI platforms (e.g., ChatGPT, Claude, Gemini, Copilot) will be able to generate campaign content perceived as relevant to the target audience.

H2. Utilization of GenAI for various Campaign activities:

The diverse use of GenAI across campaign activities will positively affect the customer experience.

H3. Content Relevance:

Higher levels of GenAI usage by marketers will be positively associated with greater perceived effectiveness in tailoring campaign content to target audiences during decision-making.

H4. Cross-Departmental Integration:

Higher levels of GenAI usage will be positively associated with greater perceived integration between marketing and other departments involved in strategic processes.

Figure 1 illustrates the conceptual model of these hypotheses:

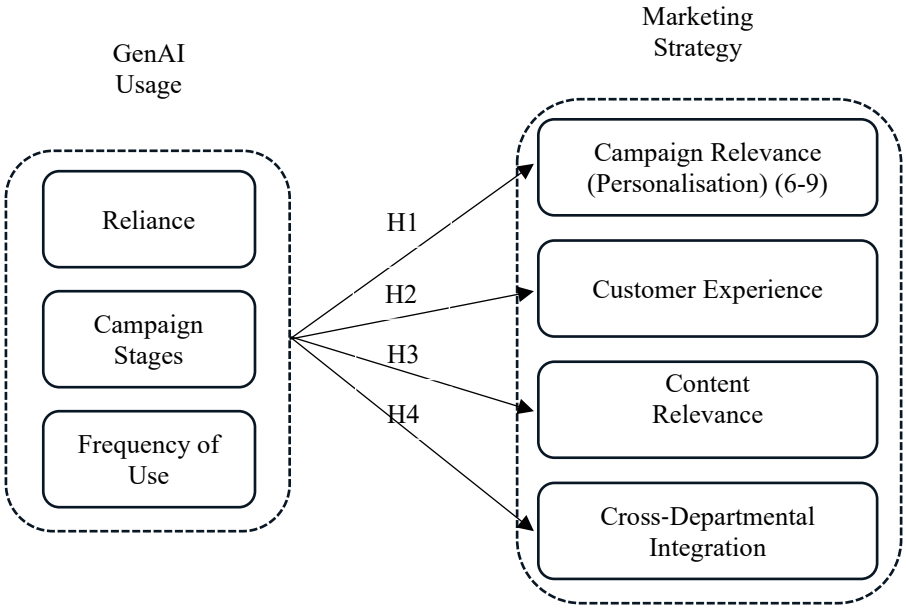


Figure 1. Conceptual model of study

III. METHODOLOGY

This study employed a correlational research design integrating two complementary methods: a survey of marketing professionals and a benchmark analysis comparing leading Generative AI (GenAI) tools. This mixed-data-source design allowed the examination of the relationships described in the research questions using two distinct quantitative sources.

1. Survey Method

The survey served as the primary data collection tool. It was chosen for its ability to gather quantifiable insights from a relevant population of marketing professionals regarding their use of GenAI and their perceptions of its influence on marketing strategy. All items were adapted from established literature on AI and marketing, ensuring conceptual validity and alignment with prior empirical work. The instrument consisted mainly of closed-ended Likert-scale questions, allowing for consistent measurement and statistical analysis. The survey included: Demographic items (e.g., role, experience level) to ensure participant relevance; GenAI usage items assessing the frequency and breadth of tool use; and perception items addressing campaign relevance, customer experience, content relevance, and interdepartmental collaboration. To ensure a common understanding of key concepts, the survey included a concise definition of GenAI at the beginning of the questionnaire. Items from source articles that were irrelevant to the present research context were intentionally excluded.

Sampling and Participants: A non-probability sampling strategy combining convenience and purposive sampling was used to reach marketing professionals with demonstrable exposure to GenAI. The survey was disseminated via professional digital communities on LinkedIn and Facebook, as well as through personal networks. Respondents were screened to confirm relevant experience.

The final sample comprised 103 marketing practitioners. Respondents were generally young, had moderate marketing experience, and most held marketing or sales positions. Importantly, all respondents reported using GenAI tools, and usage levels were relatively high. The survey questions are provided in Appendix A. Table 1 lists the variables and the number of items for each.

Table 1. Questionnaire items according to constructs

Variable	# of Statements
Reliance on GenAI	4
Campaign Stages	4
Frequency of Use	5
Campaign Relevance	4
Customer Experience	2
Content Relevancy	3
Cross-Departmental Integration	2
Total	24

Data Collection: Data were collected via Google Forms and distributed through professional messaging groups and social media channels. Participation was voluntary and anonymous, and respondents were informed that the data would be used solely for academic purposes.

Data Analysis: Given the ordinal nature of the Likert-scale items and the non-parametric approach, the primary statistical procedure was Spearman's rank-order correlation, used to examine associations between GenAI platform or usage level, and (a) campaign relevance, (b) customer experience, (c) content relevance, and (d) Cross-departmental integration.

Evaluation Metrics

Reliance on GenAI - The Reliance on GenAI was measured using a 5-level scale based on [8], an item from the information-seeking variable, replicated for each AI tool.

Campaign Stages - The extent of AI use across the campaign stages was measured on a 5-point scale, and the stages align with [22].

Frequency of Use - Frequency of use was measured on a 5-level scale, based on [23], and replicated for each AI tool.

Campaign Relevance - The campaign relevance was measured on a 5-point scale using the [8] questionnaire.

Customer Experience - The customer experience was measured on a 5-level scale based on the 8 customer experience variables.

Content Relevance - Content relevance was measured on a 5-level scale based on [24] Content Generation and Personalization variable.

Cross-Departmental Integration - The cross-departmental integration was measured using a 5-level scale based on [25], items 2 and 6.

2. Benchmark Method

The benchmark analysis complemented the survey by providing an objective comparison of GenAI platforms. Whereas the survey captures marketers' perceptions, the benchmark evaluates actual system performance in a

controlled task, aligned with Research Question 1, which focuses on the extent to which leading GenAI platforms generate relevant marketing campaigns. The benchmark addresses [8] work and consists of three main metrics: Personalisation, comprised of alignment with audience preferences, identification of explicit and implicit needs, appropriateness of spokesperson or influencer, and fit between channel choice and content style, Content Optimisation, comprised of creativity and emotional resonance, quality and clarity of messaging, and preference-based content refinement, and Purchase Intent Uplift, comprised of persuasiveness, confidence-building in decision-making, and likelihood of moving audiences toward conversion. The sub-components' metrics scores were aggregated to provide comparative performance indicators for each platform. Each metric contained multiple sub-components scored using a weighted rating system. Scores were then aggregated to provide comparative performance indicators for each platform.

Benchmark Setup: The benchmark experiment evaluated the performance of four prominent GenAI platforms (ChatGPT, Claude, Gemini, and Copilot) [26]. Each platform received an identical prompt to develop a digital campaign strategy and sample creative content for a new Coca-Cola digital initiative targeting a specified audience segment. To ensure methodological objectivity and eliminate human evaluator bias, the scoring procedure followed an automated multi-agent peer-evaluation framework modeled after [26]. After generating the respective campaign assets, the outputs from all four platforms were fed back into each platform. Using an identical evaluation prompt based on explicit scoring rubrics adapted from [8], each GenAI tool served as an autonomous evaluator, grading its own performance and the performance of the other three systems on a 1–5 scale. The final scores were assigned based on the metrics: Personalization, Content Optimization, and Purchase Intent Uplift. Each main metric comprised several sub-metrics, and scores were aggregated for each platform accordingly.

The research team rigorously verified the consistency of evaluations and the absence of systemic bias. A detailed cross-examination of the individual sub-metric scoring matrices was conducted to audit model reliability. The data confirmed stable evaluative behavior across the systems, with no single platform exhibiting anomalous grading patterns or competitive bias, such as systematically underrating rival models. By integrating both survey-based and benchmark-based methods, the study captures: Perceptions of GenAI's impact from active marketing professionals, and objective evidence of GenAI performance across platforms. This dual approach strengthens the study's ability to answer the research questions and provides a richer understanding of GenAI's role in marketing strategy. The survey questions are provided in Appendix B.

IV. RESULTS

1. Survey Results

Demographic Findings of Survey Participants

Age Distribution: The survey participants were divided into four age ranges. The majority of participants fell within the 20-30 age range (approximately 43%), with a significant portion also in the 31-40 range (approximately 28%). The 41-50 age range accounted for about 18% of participants, while the 51+ age range was represented by a smaller group (approximately 11%), as described below (See Table 2).

Table 2. Age distribution.

Age	N	%
20-30 years	44	42.7%
31-40 years	29	28.2%
41-50 years	19	18.4%

Above 51 years	11	10.7%

Years of Experience Distribution: Participants were grouped by years of experience in the field. The largest group consisted of individuals with 0-5 years of experience (approximately 38%), followed by a significant portion with 11-20 years of experience (approximately 29%). Respondents with 6-10 years of experience accounted for approximately 23% of the sample, while the smallest group was those with over 20 years of experience (approximately 10%), as described below (See Table 3).

Table 3. Experience distribution.

Experience	N	%
0-5 years	39	37.9%
6-10 years	24	23.3%
11-20 years	30	29.1%
Above 21 years	10	9.7%

Role Distribution: Participants were categorized into six distinct organizational roles. The majority of respondents held the position of Marketing Manager (approximately 41%), with a substantial portion employed in marketing (approximately 66%) and Sales (approximately 20%). Senior marketing professionals comprised about 16% of the sample, followed by Customer Relationship Managers (approximately 13%) and Market Researchers (approximately 10%). A smaller subset was represented in the Sales Manager role (approximately 1%) as described below (See Table 4).

Table 4. Positions' distribution.

Role	N	%
Customer Relationship Management	13	12.6%
Market Researcher	10	9.7%
Marketing Manager	42	40.8%
Sales	21	20.4^
Sales Manager	1	1.0%
Senior Marketing Professional	16	15.5%

Table 5 displays the model's item scores.

Table 5. Descriptive Statistics for Model Variables

		Items	M	STD
Reliance on	RG1 (Q10.1)	I rely on Claude to obtain relevant information concerning the target audience	3.37	0.12

	RG2 (Q10.2)	I rely on ChatGPT to obtain relevant information concerning the target audience	3.98	0.09
	RG3 (Q10.3)	I rely on Gemini to obtain relevant information concerning the target audience	3.71	0.09
	RG4 (Q10.4)	I rely on CoPilot to obtain relevant information concerning the target audience	3.21	0.11
Campaign Stages	CS1 (Q11.1)	To what extent do you use GenAI to target Audience Segmentation	3.51	0.10
	CS2 (Q11.2)	To what extent do you use GenAI for marketing content creation	4.17	0.09
	CS3 (Q11.3)	To what extent do you use GenAI for publishing and distributing on channels	3.04	0.11
	CS4 (Q11.4)	To what extent do you use GenAI for measuring performance	3.56	0.10
Frequency of Use	FU1 (Q5.1)	How often do you use ChatGPT	4.12	0.73
	FU2 (Q5.2)	How often do you use Claude	3.18	1.36
	FU3 (Q5.3)	How often do you, Gemini	3.45	1.21
	FU4 (Q5.4)	How often do you, CoPilot	2.90	1.25
	FU5 (Q5.5)	How often do you	2.63	1.16
Campaign Relevance	CaR1 (Q6)	Advertisements tailored by AI are very aligned with what I like	4.22	0.06
	CaR2 (Q7)	Using AI personalization simplifies the process of uncovering the items I am looking for	4.03	0.07
	CaR3 (Q8)	With AI, I can discover more accurate and meaningful information about products	4.12	0.08
	CaR4 (Q9)	AI improves my engagement with brands and makes it easy for me	3.17	0.13
Customer Experience	CE1 (Q12)	AI improves my overall interaction with brands	3.52	0.11
	CE2 (Q13)	I experience increased happiness with firms that utilize AI to enrich my experience	3.97	0.09
Content Rele	CoR1 (Q14)	Personalization of content has improved with the use of Generative AI	4.15	0.08

	CoR2 (Q15)	Generative AI helps in understanding audience preferences better for content creation	4.04	0.08
	CoR3 (Q16)	Generative AI enables us to create more relevant and personalized content	4.16	0.08
Cross-Departmental Integration	DI1 (Q17)	I communicate freely with faculty of other departments regardless of age and position	4.49	0.08
	DI2 (Q18)	I have informal contacts with other departments	4.50	0.09

2. RQ1: GenAI Platform and Campaign Relevance

Participants were asked to rate the statement "I rely on GenAI tools to obtain relevant information concerning the target audience" on a 5-point Likert scale. We examined the relationship between the GenAI platform respondents rated highest and their perceptions of the relevance of GenAI-generated campaign content to the target audience. Campaign relevance was measured using a composite score based on four items assessing personalization, preference alignment, content accuracy, and audience engagement [8].

A Spearman's rank-order correlation indicated a weak but statistically significant positive relationship between the preferred GenAI platform and perceived campaign relevance ($r = 0.35$, $p < .001$, $N = 103$), as presented in Table 6 below.

Table 6. Correlations between GenAI usage and selected outcomes (Survey data analysis)

Relationship (Variables)	Spearman's ρ	*p*-value	Interpretation
GenAI Platform vs Campaign Relevance (RQ1)	0.35	$p < .001$***	Weak-to-moderate positive correlation
GenAI Usage vs Content Relevance (RQ3)	0.32	$p = .001$**	Weak-to-moderate positive correlation
GenAI Usage vs Cross-Dept. Integration (RQ4)	0.42	$p < .001$***	Moderate positive correlation

***:p-value<0.01*

****:p-value<0.001*

Note: N = 103 for all correlations. "GenAI Usage" refers to the self-reported extent of using GenAI tools in marketing work.

The findings for RQ1 imply that marketers generally find GenAI useful for creating relevant campaigns, and align with previous literature, e.g., [6]. The Weak-to-moderate strength of correlation suggests that while GenAI use is broadly beneficial, other factors (such as the marketer's strategy or execution) may also affect campaign relevance. Additionally, this result motivated a deeper examination (through the benchmark study) of how specific platforms (ChatGPT, Claude, Gemini, Copilot) might differ in contributing to campaign relevance.

3. RQ2: AI Impact Across Campaign Stages

RQ2 addressed the effect of diverse GenAI use on customer experience. To address this, the survey asked respondents to rate the extent of GenAI's impact at several campaign stages (audience segmentation, content

creation, distribution, and performance measurement). First, we analyzed the correlation between aggregated GenAI use across all marketing activities and user experience, and then performed separate analyses for each of the four marketing activities using Spearman's rank correlation. The effect of the aggregated activities was found to be positive, r = 0.51, p < .001, N = 103, and so does each of the effects of the various activities (See Table 7)

Table 7. Relationship between GenAI usage and impact at different campaign stages *(Spearman correlations, N = 103)*

Campaign Stage	Spearman r	*p*-value	Interpretation
The four campaign stages	0.51	$p < .001$***	Moderate positive correlation
Audience Segmentation	0.42	$p < .001$***	Moderate positive correlation
Content Creation	0.29	$p < .01$**	Weak positive correlation
Content Distribution	0.41	$p < .001$***	Moderate positive correlation
Performance Measurement	0.39	$p < .001$***	Weak-to-moderate positive correlation

***:p-value<0.01*

****:p-value<0.001*

As shown**,** all activities exhibit a positive correlation between GenAI usage and perceived impact, with all correlations statistically significant and differing in strength. Of the four campaign stages, the strongest observed relationship was for Audience Segmentation ($r = 0.42$, $p < .001$), indicating that higher use of GenAI is fairly strongly associated with greater improvements in segmentation tasks.

These findings support Hypothesis 2, which posited that GenAI use across diverse campaign activities would affect customer experience, and also reveal that GenAI use in separate campaign activities would do so as well.

Delving into the details, survey respondents indicated that GenAI is most commonly used in the content creation stage of campaigns, which is unsurprising given tools like ChatGPT's ability to generate copy or creative ideas. However, the statistical analysis reveals that the greatest benefit of GenAI occurs at the segmentation stage.

Overall, the results for RQ2 indicate that GenAI positively influences all examined activities in a marketing campaign, particularly in audience segmentation and customer experience. This aligns with studies such as [12], which highlight AI's value in advanced segmentation and suggest that organizations could benefit from employing GenAI tools more extensively beyond content generation, including in the planning and analysis phases.

4. RQ3: GenAI Usage for Content Adaptation and Decision-Making

RQ3 focused on how extensively marketers use GenAI to tailor campaign content to the target audience during campaign decision-making. To measure this, the survey included an index of GenAI usage intensity (combining the frequency of GenAI tool use and the number of GenAI platforms used) and an index of content adaptation effectiveness (combining ratings of GenAI's contribution to creating relevant, personalized campaign content and to improving decision quality). A Spearman correlation was computed between GenAI usage intensity and content adaptation effectiveness. The analysis yielded a significant positive correlation (r = 0.32, N = 103, p = .001), indicating that greater GenAI use is associated with better tailoring of campaign content to audiences and improved campaign decision-making. This statistically significant relationship leads us to reject the null hypothesis and supports Hypothesis 3.

In practice, this finding means that marketers who more frequently or more broadly employ GenAI tools tend to report greater success in customizing content for their target audiences and feel that GenAI aids their campaign-related decision-making. These findings are consistent with prior work [27] showing that GenAI enables superior content selection, targeting, and dynamic optimization, as well as with studies confirming that AI-personalization boosts relevance and ad performance [14].

Usage levels were high (M = 4.32, between "often" and "always") and showed little variance, signaling broad adoption. Perceived effectiveness for personalized content was likewise high (M = 4.11 on a 5-point scale). Notably, despite this uniform uptake, a positive correlation suggests that heavier users, those leveraging multiple GenAI tools, derive incremental benefits in content relevance and decision speed. Taken together, these results confirm that GenAI is now deeply embedded in routine marketing work, and that deeper engagement corresponds to stronger personalization capabilities and more agile decision-making.RQ4: GenAI Usage and Cross-Departmental Integration

5. RQ4: GenAI Usage for Cross-Department Integration

RQ4 examined whether greater use of GenAI tools in marketing is associated with higher levels of integration between marketing teams and other departments (such as IT or data science) within the strategic marketing process. To explore this, respondents rated their organization's level of interdepartmental integration (via two items on open communication and collaboration with other departments). These items were correlated with the GenAI usage intensity index described above. The result was a positive and significant ($p < .001$, $N = 103$, $r = .42$) relationship between GenAI usage and cross-department integration, indicating that the more marketers use GenAI, the better they integrate with other departments. Thus, Hypothesis 4, which posited a positive link between GenAI use and interdepartmental integration, was supported by the survey data.

Interestingly, respondents generally reported a high level of cross-department integration in their organizations (mean ~4.49 on a 5-point scale, close to "to a very great extent"), with relatively low variability (SD ~0.08). Many marketers reported strong communication and cooperation with other departments (e.g., IT and data analytics).

This finding aligns with the literature, as [21] suggests that GenAI platforms serve as a semantic layer linking data, knowledge, and people, thereby theoretically requiring greater cross-functional synchronization.

In summary, the survey results for RQ4 indicate a significant relationship between GenAI usage and the integration of GenAI into marketing processes.

Finally, to synthesize the survey findings: the participant profile and responses suggest marketers today are a relatively young cohort with moderate experience, almost all of whom use GenAI frequently in their work. They view GenAI as a significant tool that helps create more relevant campaigns, especially in personalization, information accuracy, and content generation, and that its use in campaign activities positively affects the customer experience. GenAI is also seen as having a positive impact on content adaptation and decision-making, as well as on collaboration with other departments.

V. BENCHMARK RESULTS

This section presents the key findings from the benchmark experiment, which compared the performance of four prominent GenAI platforms (ChatGPT, Claude, Gemini, Copilot) in generating a marketing campaign for a global brand (Coca-Cola). The benchmark aimed to address a focused RQ1 from a different angle: by objectively evaluating how well each GenAI tool produces relevant and effective marketing content for the target audience.

Personalization Metric – Results: for the overall personalization score—which combines all sub-metrics related to tailoring content to the audience—Claude and Gemini achieved the highest scores, indicating excellent performance. Claude posted the top weighted score of 1.54, narrowly edging out Gemini's 1.52 on the personalization index. In contrast, ChatGPT recorded the lowest overall score at 1.21, reflecting comparatively weaker personalization, while Copilot fell between these extremes.

At the sub-metric level, Gemini led in the analysis of hidden and explicit needs with a score of 0.37, whereas ChatGPT lagged at 0.29. Gemini also topped the influencer/persona selection sub-metric with a score of 0.35, while ChatGPT ranked lowest at 0.29. For alignment of channel and content style, Gemini and Claude tied for the best performance (each scoring 0.42), with ChatGPT trailing at 0.33. Similarly, for matching content to audience preferences, Gemini and Claude again tied at 0.40, compared to ChatGPT's 0.29.

These results illustrate that Gemini and Claude consistently outperformed ChatGPT on personalization-related measures. Gemini showed strength in deeply understanding customer needs and choosing appropriate content spokespeople and channels, whereas Claude was equally strong in ensuring content style alignment and preference matching. ChatGPT, on the other hand, tended to produce more generic content that was less finely tuned to the specifics of the target segment, particularly for a well-known brand like Coca-Cola. Copilot often performed comparably to Gemini and Claude on certain sub-metrics (for instance, Copilot tied with them on at least one sub-metric, scoring 0.42 in "relevance and personal utility" of content).

Content Optimization Metric – Results: In the content optimization metric (evaluating creativity, emotional resonance, and content quality), Claude ranked first. Claude achieved the highest weighted score of 1.44 in content optimization. ChatGPT again was the lowest, with a score of 1.06 on this metric. Notably, ChatGPT not only had the lowest composite score here but also the lowest scores on nearly all the sub-metrics under content optimization: for example, in the sub-metric of *"increasing interest and excitement with the content*, " ChatGPT led with a score of 0.53, indicating it was best at producing content that could emotionally engage the audience. In *preference-based content optimization*, Gemini and Claude tied with high scores (0.5 each), whereas ChatGPT scored only 0.31 – a substantial gap.

These comparisons indicate that Claude is particularly strong at generating creative content that resonates emotionally and at optimizing content based on audience preferences. Gemini was a close competitor in many of these aspects as well, showing strength in data-driven content adjustments. ChatGPT's content tended to be more generic and less optimized for specific emotional triggers or preferences, which is why it scored poorly on these nuanced measures. Copilot's performance was mixed; in some respects, it matched the top tools (for instance, Copilot tied with Gemini and Claude on the "personal relevance and utility" sub-metric at 0.42 and was close behind in other sub-metrics).

Purchase Intent Uplift Metric – Results: Gemini and Claude were the leaders for the purchase intent uplift metric. Gemini achieved a slightly higher score (1.33) compared to Claude (1.19) on the overall purchase intent index, with Copilot close behind (1.14). ChatGPT was again lower than the others on sub-metrics related to purchase intent: on *the persuasiveness of advertisements, Gemini and Claude tied for the top score* of 0.44. On *boosting confidence in purchase decisions*, Gemini was highest at 0.47, with Claude and Copilot next (0.39 each). On *leading exposure to purchase*, Gemini led with 0.42, followed by Claude at 0.36, whereas ChatGPT was lowest at 0.28 on this sub-metric.

These results suggest Gemini has a slight edge in driving purchase-related outcomes, possibly due to its strength in data analysis and understanding consumer decision factors (as evidenced by its superior performance in needs analysis and confidence boosting). Claude is not far behind, however, and actually surpasses Gemini in other areas (such as *creative content generation*). Copilot showed itself to be a solid performer in supporting purchase intent, close to Gemini and Claude, which may reflect its design as an assistive model for tasks.

Key Benchmark Findings: An analysis of the overall performance rankings reveals a clear quality gap among the GenAI platforms: Gemini achieved the highest aggregate performance score across all metrics (total score ≈ 4.27 out of a possible range), making it the top all-around performer in our benchmark. Gemini particularly

excelled in tasks related to purchase intent and deep personalization. It was the most effective tool for understanding consumer data and translating that understanding into stronger purchase intent. Specifically, Gemini outperformed Claude on sub-metrics such as boosting purchase decision confidence and providing targeted exposure that drives conversion, underscoring Gemini's strength as *"the most data-precise tool"* in the test. Claude was a close competitor, outperforming Gemini on the content optimization metric (Claude's 1.44 vs Gemini's 1.42) and the overall personalization metric (Claude's 1.54 vs Gemini's 1.52). Claude also scored highest in generating emotionally engaging content (e.g., *raising content interest/excitement*). These outcomes indicate that Claude is the best tool for producing creative, emotionally resonant marketing content, even though Gemini might have a slight edge in data-driven precision. One specific insight is that Claude's content tends to evoke emotion effectively, a crucial aspect of marketing content, meaning Claude can make marketing messages more compelling on a human level. In contrast, Gemini might be more pinpoint in targeting. ChatGPT consistently received the lowest scores across most metrics and sub-metrics. This suggests a performance gap, in which ChatGPT's output was more generic and less finely attuned to the specific marketing context we tested. ChatGPT appeared to struggle, especially with preference-based optimization (score 0.31 vs 0.5 for Gemini/Claude) and other personalization subtleties, suggesting it may lack advanced domain fine-tuning or the latest model capabilities that the other tools have. It tended to produce "one-size-fits-all" content for a large brand like Coca-Cola, whereas the newer models (Gemini, Claude) dug deeper into nuance. This highlights the "digital gap" with ChatGPT in this scenario: it may be a powerful general-purpose tool, but for sophisticated marketing tasks, it lags behind specialized or newer GenAI models. Copilot performed respectably, often ranking in the middle. It tied with the leaders on some personalization sub-metrics and was only slightly behind on purchase-intent measures. Copilot did not clearly lead any single category, but it proved to be a well-rounded tool.

In summary, the benchmark findings show that marketing effectiveness in the GenAI era depends on choosing the right tool for the specific task at hand. Gemini stands out for data-centric tasks and driving conversions. At the same time, Claude excels at creating engaging content and fostering emotional engagement. At the same time, ChatGPT may not be the optimal choice when high levels of personalization or fine-tuned optimization are required for a campaign.

VI. DISCUSSION

The overarching goal of this study was to examine how marketing strategies adapt to Generative AI (GenAI). Converging survey and benchmark data indicate that GenAI drives significant structural changes across marketing strategy.

Regarding campaign relevance to the target audience, both methodologies confirm that GenAI enhances campaign effectiveness. While the survey demonstrates a baseline positive correlation across platforms, the benchmark adds nuance by showing that specific platform architecture matters. While any GenAI can improve personalization, corroborating research by [6] on AI-driven personalized customer journeys, our results stress that matching the specific platform to the audience profile is vital for optimal message-audience fit.

For the campaign process stages, we hypothesized that GenAI's most profound impact would occur during audience segmentation. The survey partially supported this assumption: GenAI usage correlated most strongly with advanced audience segmentation, yet it showed lower relative impact in content creation, where it is paradoxically most frequently deployed. This confirms that GenAI's analytical capabilities shine brightest in upstream strategic phases, as outlined in earlier theoretical frameworks by [10]. The gap between content generation execution and segmentation impact reveals a critical strategic underutilization.

In terms of content personalization and decision support, extensive GenAI integration yields superior alignment with the content audience and accelerated, data-informed decisions. This supports the findings of [27] and [8], who noted that GenAI fundamentally reshapes how content is delivered. The moderate correlation ($\rho = 0.32$)

confirms that while GenAI is a powerful computational aid that makes marketing iterative and agile, human expertise and creative oversight remain foundational.

Furthermore, survey results confirmed a positive correlation between GenAI adoption and cross-functional integration across marketing, IT, and data analytics departments. Because modern digital marketing relies on Management Information Systems as an organizational backbone (as [19] emphasized), GenAI functions as a connective semantic layer, in line with the findings of [21]. However, our data suggest that the depth of this integration is ultimately bound to organizational culture and leadership structure, mirroring observations in [17].

Combining empirical approaches reveals two key strategic insights regarding methodological integration. First, the survey indicates a general benefit of GenAI adoption, but the benchmark shows that performance variance between models is non-trivial. Marketers must deploy platforms based on discrete task strengths, such as leveraging Claude for emotional resonance and Gemini for structured data analysis. Second, a clear execution-strategy gap emerged. Marketers heavily favor GenAI for downstream content execution, yet capabilities benchmarked for decoding consumer needs are associated with higher purchase intent. Marketing strategy must evolve to embed GenAI into front-end strategy formulation rather than treating it merely as a high-speed copywriter.

The findings carry several implications for theory. This study enriches the literature on emerging technology adoption by identifying an empirical gap between utilization frequency and latent strategic potential. While current frameworks assume linear adoption paths, our findings suggest that organizations gravitate toward high-visibility, low-complexity tasks such as content generation, while underutilizing high-impact analytical capabilities, such as segmentation. This underscores the need for new theoretical models to explain how organizational learning-curve dynamics shift as functional familiarity with AI matures.

Practically, the study offers data-driven insights for marketing managers. Marketing managers should shift GenAI from simple tactical execution to predictive modeling, customer clustering, and strategic scenario testing. Additionally, organizations must move away from a one-size-fits-all software procurement model. Teams should build a diversified portfolio of GenAI assets, choosing specific large language models based on empirical campaign objectives. Finally, rather than training staff purely on prompt engineering for text or image generation, training frameworks should focus on data synthesis and strategic interpretation using AI-driven analytical tools.

This study provides a critical empirical baseline for the Israeli ecosystem. The data demonstrates that local marketing professionals are successfully keeping pace with international AI integration benchmarks established between 2023 and 2025. In conclusion, GenAI is a transformative yet unevenly leveraged force in marketing strategy. Maximizing its value requires organizations to actively broaden their deployment parameters, moving past routine execution to reshape the foundational layers of strategic marketing management.

VII. CONCLUSION

1. Summary of Findings

This research reveals that while GenAI significantly enhances marketing strategies, its effectiveness varies considerably across functional dimensions. Although GenAI demonstrably improves campaign relevance and content personalization, the quality of the outcomes is heavily contingent on the specific tool selected. A notable discrepancy emerged between current usage patterns and true potential: while practitioners predominantly apply GenAI to content creation, its most substantial impact was found in the strategic stage of audience segmentation. Furthermore, the results confirm a positive relationship between GenAI usage and decision-making efficacy.

2. Research Statement

These findings collectively underscore that realizing strategic value from GenAI demands more than mere adoption—it requires targeted, intentional implementation. Marketers should deliberately align platform selection with specific campaign objectives to maximize audience relevance. They should expand GenAI deployment beyond content generation into higher-impact strategic functions such as segmentation and real-time data analysis, where it can significantly enhance agility and precision. Critically, organizations must recognize GenAI as a facilitator rather than a cure for structural fragmentation; its cross-functional benefits depend on pre-existing collaborative cultures and robust human processes. Ultimately, GenAI's full potential is unlocked only when it is deeply integrated into decision workflows and supported by an organizational environment that values both technological and interpersonal coordination.

3. Limitations and Future Research Directions

This study has several limitations that open avenues for future inquiry. First, the sample comprised approximately 100 marketing professionals, primarily based in Israel, and the measures relied on subjective self-reports rather than objective performance data. However, we attempted to mitigate this by including a benchmark experiment. Future research should expand to larger, more diverse international samples across industries and countries, and complement perceptual data with objective metrics, such as A/B test results comparing AI-generated versus non-AI content or longitudinal KPI trends following GenAI adoption.

Second, the rapidly evolving nature of GenAI presents a temporal limitation, as our benchmarks reflect tool capabilities at a single point in time; subsequent model updates or new market entrants could shift the competitive rankings we observed. Researchers should periodically re-evaluate platform performance as updates are released, and practitioners must remain agile in reassessing which tools best suit their strategic needs.

Additionally, extending benchmarks to encompass a wider array of GenAI platforms, marketing tasks, and brand contexts. including B2B scenarios and diverse industries. would test the generalizability of our findings. Controlled field experiments that isolate GenAI's specific effects relative to traditional campaigns are also warranted. More broadly, as GenAI becomes further embedded in marketing practice, ongoing inquiry must address emerging challenges surrounding ethics, the evolving role of human creativity, and the new skill sets marketers will require. Our work provides a current snapshot to anchor these critical, evolving discussions.

ACKNOWLEDGEMENTS

The author gratefully acknowledges May Zehavi, Yuval Sidi, and Koreen Trif, whose original student seminar project provided the foundational research framework and data collection that inspired this paper. Although they were not involved in the drafting or finalization of this manuscript, their initial academic contributions were vital to the realization of this study.

DECLARATION OF AI USE

During the preparation of this work, the authors used OpenAI's ChatGPT to assist with translating the original Hebrew content into English and refining the manuscript's language. The AI tool was employed to ensure clarity and fluency in academic writing. All content was then reviewed and edited by the authors to verify accuracy and coherence with the research findings. Additionally, the benchmark that addressed the four GenAI tools: ChatGPT, Claude, Gemini, and Copilot, was executed by these tools, as described in the manuscript.

REFERENCES

[1] M. Ertz and M. Kordi, "Advancing Marketing Strategy with Artificial Intelligence," in *AI in Marketing*, 1st ed., London: Routledge, 2025, pp. 22–48. doi: 10.4324/9781003468806-3.

[2] M. Potwora, O. Vdovichena, D. Semchuk, L. Lipych, and V. Saienko, "The use of artificial intelligence in marketing strategies: Automation, personalization and forecasting," *J. Manag. World*, vol. 2024, no. 2, pp. 41–49, May 2024, doi: 10.53935/jomw.v2024i2.275.

[3] A. Stancu and M. Panait, "Marketing Strategy Metamorphosis Under the Impact of Artificial Intelligence Services," *Systems*, vol. 13, no. 4, p. 227, Mar. 2025, doi: 10.3390/systems13040227.

[4] G. Wilson, O. Johnson, and W. Brown, "The Impact of Artificial Intelligence on Digital Marketing Strategies," Aug. 05, 2024, *Business, Economics, and Management*. doi: 10.20944/preprints202408.0276.v1.

[5] E. Hermann and S. Puntoni, "Generative AI in Marketing and Principles for Ethical Design and Deployment," *J. Public Policy Mark.*, vol. 44, no. 3, pp. 332–349, Jul. 2025, doi: 10.1177/07439156241309874.

[6] D. Grewal, C. B. Satornino, T. Davenport, and A. Guha, "How generative AI is shaping the future of marketing," *J. Acad. Mark. Sci.*, vol. 53, no. 3, pp. 702–722, May 2025, doi: 10.1007/s11747-024-01064-3.

[7] J. Hartmann, Y. Exner, and S. Domdey, "The power of generative marketing: Can generative AI reach human-level visual marketing content?" *SSRN Electron. J.*, 2023, doi: 10.2139/ssrn.4597899.

[8] H. Beyari and T. Hashem, "The Role of Artificial Intelligence in Personalizing Social Media Marketing Strategies for Enhanced Customer Experience," *Behav. Sci.*, vol. 15, no. 5, p. 700, May 2025, doi: 10.3390/bs15050700.

[9] N. Kshetri, Y. K. Dwivedi, T. H. Davenport, and N. Panteli, "Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda," *Int. J. Inf. Manag.*, vol. 75, p. 102716, Apr. 2024, doi: 10.1016/j.ijinfomgt.2023.102716.

[10] M.-H. Huang and R. T. Rust, "A strategic framework for artificial intelligence in marketing," *J. Acad. Mark. Sci.*, vol. 49, no. 1, pp. 30–50, Jan. 2021, doi: 10.1007/s11747-020-00749-9.

[11] C. Ziakis and M. Vlachopoulou, "Artificial Intelligence in Digital Marketing: Insights from a Comprehensive Review," *Information*, vol. 14, no. 12, p. 664, Dec. 2023, doi: 10.3390/info14120664.

[12] F. M. Talaat, A. Aljadani, B. Alharthi, M. A. Farsi, M. Badawy, and M. Elhosseini, "A Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing," *Mathematics*, vol. 11, no. 18, p. 3930, Sep. 2023, doi: 10.3390/math11183930.

[13] K. Abrokwah-Larbi and Y. Awuku-Larbi, "The impact of artificial intelligence in marketing on the performance of business organizations: evidence from SMEs in an emerging economy," *J. Entrep. Emerg. Econ.*, vol. 16, no. 4, pp. 1090–1117, Jun. 2024, doi: 10.1108/JEEE-07-2022-0207.

[14] T. Li, "AI Empowered Personalized Digital Advertising Marketing," *Front. Bus. Econ. Manag.*, vol. 15, no. 2, pp. 439–442, May 2024, doi: 10.54097/k0wjgk28.

[15] M. B. Pinto, J. K. Pinto, and J. E. Prescott, "Antecedents and Consequences of Project Team Cross-Functional Cooperation," *Manag. Sci.*, vol. 39, no. 10, pp. 1281–1297, Oct. 1993, doi: 10.1287/mnsc.39.10.1281.

[16] A. E. Ellinger, S. B. Keller, and A. D. Ellinger, "Developing Interdepartmental Integration: An Evaluation of Three Strategic Approaches for Performance Improvement," *Perform. Improv. Q.*, vol. 13, no. 3, pp. 41–59, Oct. 2008, doi: 10.1111/j.1937-8327.2000.tb00175.x.

[17] Rita Uchenna Attah, Baalah Matthew Patrick Garba, Ifechukwu Gil-Ozoudeh, and Obinna Iwuanyanwu, "Cross-functional team dynamics in technology management: a comprehensive review of efficiency and innovation enhancement," *Eng. Sci. Technol. J.*, vol. 5, no. 12, pp. 3248–3265, Dec. 2024, doi: 10.51594/estj.v5i12.1756.

[18] L. Zhou, S. Yan, Z. Li, L. Song, J. Ma, and Y. Yang, "Practice of Inter-departmental Collaboration Systems under a Tiered Data Sharing Framework and Integrated Application of Large Language Models," in *2025 40th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Zhengzhou, China: IEEE, May 2025, pp. 2283–2288. doi: 10.1109/YAC66630.2025.11150052.

[19] M. Lishmah Dominic, P. S. Venkateswaran, L. T. Reddi, S. Rangineni, R. Regin, and S. S. Rajest, "The Synergy of Management Information Systems and Predictive Analytics for Marketing," in *Advances in Business Information Systems and Analytics*, S. Singh, S. S. Rajest, S. Hadoussa, A. J. Obaid, and R. Regin, Eds., IGI Global, 2023, pp. 49–63. doi: 10.4018/979-8-3693-2193-5.ch004.

[20] S. Bergen, "Integrating Artificial Intelligence to Enhance Business Processes: Strategies for Effective Decision-Making, Predictive Analytics, and Process Automation Amidst Technical and Organizational Challenges," Tilburg, Netherlands.

[21] L. Reinkemeyer, "Business Perspective: The Future of Intelligent Business Process Execution," in *Process Intelligence in Action*, L. Reinkemeyer, Ed., Cham: Springer Nature Switzerland, 2024, pp. 233–245. doi: 10.1007/978-3-031-61343-2_23.

[22] Kamaal Allil and Adil Zia, "Digital Age Content Marketing: Strategies and Trends," *Micro Macro Mark.*, no. 3, pp. 611–638, 2024, doi: 10.1431/113076.

[23] A. Chatzopoulos, P. Zacharia, and A. Kantaros, "Generative AI in Higher Education: A Large-Scale Study of Student Usage Patterns, Applications and Motivations," *Appl. Sci.*, vol. 16, no. 12, p. 5972, Jun. 2026, doi: 10.3390/app16125972.

[24] V. Soni, "Adopting generative AI in digital marketing campaigns: An empirical study of drivers and barriers," *Sage Sci. Rev. Appl. Mach. Learn.*, vol. 6, no. 8, pp. 1–15, Aug. 2023.

[25] A. Mubashra and M. A. A. Saleemi, "An Analysis of Inter-Departmental Integration towards Organizational Development," *Hum. Nat. J. Soc. Sci.*, vol. 4, no. 2, pp. 158–171, Jun. 2023, doi: 10.71016/hnjss/n87m7232.

[26] R. Gafni and Y. Levy, "Comparing GenAI platforms on cybersecurity management task performances," *Inf. Comput. Secur.*, vol. 34, no. 1, pp. 21–46, Jan. 2026, doi: 10.1108/ICS-04-2025-0164.

[27] H.-L. Chan and T.-M. Choi, "Using generative artificial intelligence (GenAI) in marketing: Development and practices," *J. Bus. Res.*, vol. 191, p. 115276, mar. 2025, doi: 10.1016/j.jbusres.2025.115276.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Tabletop Exercise Framework for Cross-Border Digital Health Incident Response

Niko Haanpää
Laurea University of Applied Sciences
Espoo, Finland
niko.haanpaa@student.laurea.fi

Teemu Koski
Laurea University of Applied Sciences
Espoo, Finland
teemu.koski@student.laurea.fi

Tommi Elg
Laurea University of Applied Sciences
Espoo, Finland
tommi.elg@student.laurea.fi

Ville Kivistö
Laurea University of Applied Sciences
Espoo, Finland
ville.kivisto@student.laurea.fi

Jyri Rajamäki
Laurea University of Applied Sciences
Espoo, Finland
jyri.rajamaki@laurea.fi

Abstract
The increasing interoperability of digital health systems creates cybersecurity challenges that extend across organizational and national boundaries. As healthcare organizations exchange data through interconnected infrastructures, incidents may rapidly evolve into multi-stakeholder events requiring coordinated response, communication, and decision-making. This work-in-progress paper presents the design of a reusable tabletop exercise toolkit to improve incident-response preparedness in cross-border digital health environments. Developed using a Design Science Research approach, the toolkit consists of a Facilitator Guide and a Participant Pack supporting a structured 60-90-minute exercise simulating a cybersecurity incident affecting digital health interoperability services. The exercise incorporates progressive scenario injects, role-based participation, an observer scorecard, and a structured hotwash process to support evaluation and organizational learning. The toolkit is designed to expose interoperability-related dependencies, strengthen stakeholder coordination, and support decision-making under uncertainty. Although formal validation has not yet been conducted, the proposed design demonstrates how scenario-based tabletop exercises can support cyber resilience in increasingly interconnected healthcare ecosystems. Future work will focus on pilot testing, usability evaluation, and assessment of preparedness improvements achieved through repeated exercises.

Keywords: cybersecurity, incident response, tabletop exercise, digital health interoperability, cyber resilience, healthcare cybersecurity, cross-border collaboration

I. INTRODUCTION

Digital health systems are becoming increasingly interconnected across organizational [1] and national boundaries [2]. As healthcare organizations exchange electronic health information through interoperable services, cybersecurity incidents may extend beyond a single organization and require coordinated response activities involving multiple stakeholders [3]. Effective incident response is therefore becoming an essential aspect of secure digital health operations [4]. Although healthcare organizations commonly utilize incident response plans, coordination across organizational and national boundaries remains challenging [5]. Traditional awareness activities and policy documents rarely provide opportunities to practice communication, decision-making, and evidence handling in realistic cross-border scenarios. As a result, weaknesses in operational preparedness may only become visible during an actual incident. This study was motivated by the need for a

practical preparedness tool that supports incident response exercises in interoperable healthcare environments. Interoperability is increasingly understood not merely as the technical exchange of data but as the combined capability of heterogeneous systems, organizations, and jurisdictions to exchange information, preserve its meaning, integrate it into operational workflows, and do so within agreed legal frameworks. Rather than developing another checklist or policy document, the paper focuses on creating a reusable tabletop exercise toolkit that enables organizations to rehearse realistic cybersecurity incidents in a structured and time-efficient manner. The objective is to provide a practical environment in which participants can explore not only technical response activities but also the organizational, communication, coordination, and accountability challenges that emerge when interoperability-dependent services are disrupted. The primary research question guiding this work is: How can a 60–90-minute tabletop exercise be structured to realistically practice cross-organizational roles, communication, decision-making, and evidence handling in cross-border digital health incident response? The working hypothesis is that a role-based and scenario-specific tabletop exercise toolkit can improve preparedness by making triage decisions, containment actions, communication processes, and evidence-handling responsibilities more visible and easier to evaluate. The objective of this work-in-progress paper is to design a reusable tabletop exercise toolkit to support cyber resilience in interoperable healthcare environments by facilitating observation and improvement of coordination, communication, and decision-making processes during cybersecurity incidents.

II. BACKGROUND AND RELATED WORK

Digital health interoperability enables healthcare organizations to exchange health information across organizational and national borders. The European Health Data Space (EHDS) seeks to establish a common framework for the secure exchange and use of electronic health data throughout the European Union. By promoting interoperability, patient mobility, and cross-border access to healthcare information, EHDS will significantly increase the interconnectedness of healthcare systems [6]. While these developments offer substantial benefits for healthcare delivery and continuity of care, they also increase the potential impact of cybersecurity incidents. A disruption affecting a single component of a cross-border ecosystem may have consequences for multiple organizations and jurisdictions. Consequently, cyber resilience and incident response preparedness are becoming critical enablers of successful EHDS implementation [3]. Cross-border digital health services create dependencies that extend beyond individual healthcare organizations [7]. Healthcare providers increasingly rely on identity and access management systems, interoperability gateways, national contact points (NCPs), centralized logging infrastructures, cloud services, vendors, and cross-organizational data flows [8]. Health data is highly sensitive, meaning that both service delivery and service disruption have significant privacy and regulatory implications. Interoperability within the European Union must be considered together with data protection requirements, governance structures, and the lawful handling of health information [9]. Consequently, incident response cannot be viewed solely as a technical recovery activity. The implementation of EHDS further increases the need for coordinated response capabilities. Healthcare providers, national authorities, system vendors, and infrastructure operators must cooperate to maintain service continuity, protect personal health information, and comply with legal requirements while responding to cybersecurity incidents [3]. A key observation underpinning this paper is that many operational dependencies remain hidden until organizations are required to respond to an incident under conditions of uncertainty. Based on the research materials, four dependency areas were identified as particularly significant: (1) Identity services [10], (2) Gateway or National Contact Point (NCP) operations [7], (3) Vendor and third-party access [11], and Cross-organizational data flows [12]. These dependencies act as practical decision bottlenecks that influence communication strategies, evidence collection, containment decisions, and service restoration activities. Tabletop exercises are widely used to support preparedness by enabling participants to discuss and evaluate their responses to realistic incident scenarios in a controlled environment [13]. However, existing exercises often focus on organizational or technical aspects of incident management rather than on the specific dependencies arising from cross-border digital health interoperability. The contribution of this work is therefore not a new theory of healthcare cybersecurity but a reusable training asset that transforms interoperability-related cybersecurity concerns into a

structured exercise format. By doing so, the paper seeks to move from general awareness of interoperability risks to practical rehearsal of incident-response behaviors in cross-border healthcare environments.

III. RESEARCH METHOD AND DESIGN APPROACH

This work was conducted using a Design Science Research (DSR) approach [14], in which the primary objective is to create and evaluate a practical artifact that addresses a real-world problem. In this study, the identified problem concerns the limited opportunities available for healthcare organizations to rehearse incident response activities in cross-border digital health environments characterized by interoperability dependencies, multiple stakeholders, and incomplete information. Rather than proposing a new theoretical model of healthcare cybersecurity, the paper focuses on developing a reusable tabletop exercise toolkit that enables realistic discussion and evaluation of cybersecurity incident response processes. The resulting artifact is intended to support preparedness, coordination, communication, and decision-making across organizational and national boundaries. The design process was guided by the assumption that cyber resilience in interoperable healthcare ecosystems depends not only on technical controls but also on organizations' ability to coordinate actions, manage dependencies, and make defensible decisions under uncertainty. Consequently, the toolkit was designed to expose operational dependencies and stimulate cross-functional discussion among technical, managerial, legal, privacy, and communication stakeholders. The motivation for this study emerged from the increasing interconnectedness of digital health services and the growing importance of cross-border healthcare information exchange. Initiatives such as the European Health Data Space (EHDS) are expected to further increase interoperability among healthcare organizations, national infrastructures, and service providers. While these developments create significant benefits for healthcare delivery, they also introduce complex cybersecurity dependencies that may span multiple organizations and jurisdictions. The dependencies that repeatedly influence incident response activities in digital health environments often remain invisible during normal operations but become critical during cybersecurity incidents. They affect evidence collection, communication, containment decisions, recovery actions, and accountability across organizational boundaries. The design challenge was therefore to create a practical exercise environment that could reproduce these coordination challenges within a limited time frame while remaining realistic and repeatable. The toolkit was developed according to four operational design principles derived from the identified problem domain and exercise objectives: (1) Rapid identification of incident ownership to ensure that responsibility and escalation paths become visible early in the exercise; (2) Visibility of interoperability-related dependencies to expose how identity services, gateway operations, suppliers, and data flows influence response activities; (3) Decision-making under uncertainty to reflect the reality that cybersecurity incidents often begin with incomplete or contradictory information. Participants are encouraged to state assumptions explicitly and justify their decisions despite uncertainty; and (4) Generation of actionable improvement outcomes so that observations made during the exercise can be converted into concrete corrective actions and organizational learning. These principles guided the structure of the exercise scenario, participant roles, evaluation mechanisms, and post-exercise activities. The artifact development resulted in a Facilitator Guide and a Participant Pack, which are described in detail in Section IV. The exercise was designed to be completed within 60–90 minutes and organized around a sequence of scenario injects that progressively increase decision pressure and organizational complexity. Each injection introduces new information that requires participants to reassess assumptions, coordinate stakeholders, and make operational decisions based on the information available at that moment. The development process also incorporated role-based participation. Six operational roles were defined to represent the principal stakeholder groups typically involved in cross-border cybersecurity incident response: (1) Incident Lead, (2) Technical Lead, (3) Gateway/NCP Owner, (4) Privacy Officer (DPO), (5) Communications Lead, and (6) Vendor Liaison. This structure was chosen to ensure that technical, legal, organizational, privacy, and supplier perspectives are represented throughout the exercise. A central objective of the toolkit is not merely to facilitate discussion but also to enable systematic observation of decision-making behaviors. For this reason, evaluation was built into the design. The exercise concludes with a structured hotwash process in which participants review the challenges encountered and identify opportunities for improvement. Required outputs include corrective actions, assigned

owners, implementation deadlines, and recommendations for future exercises or validation activities. By integrating scenario execution, observation, and structured reflection into a single process, the toolkit was designed to serve as both a preparedness assessment instrument and a mechanism for continuous organizational improvement.

IV. CROSS-BORDER INCIDENT RESPONSE TOOLKIT

The proposed artifact is a reusable tabletop exercise toolkit designed to support cybersecurity preparedness in interoperable digital health environments. The toolkit is intended to support discussion-based learning, structured decision-making, and the identification of operational dependencies that influence cross-border incident response. The architecture contains two complementary exercise components: a Facilitator Guide and a Participant Pack. Together, they provide a common scenario, predefined exercise injects, role descriptions, guidance material, and evaluation mechanisms. The design follows four operational principles: (1) Rapid identification of incident ownership; (2) Visibility of interoperability-related dependencies; (3) Decision-making under uncertainty; and (4) Generation of actionable organizational improvements following the exercise. Fig. 1 illustrates the overall architecture of the toolkit and its learning cycle. The toolkit consists of a Facilitator Guide and a Participant Pack that jointly support a scenario-driven exercise process. The Scenario Engine delivers three progressively escalating injects covering detection and triage, containment and cross-border coordination, and restoration and learning. The architecture is intended to transform interoperability-related cybersecurity concerns into a structured, repeatable, and time-efficient process that can be completed in 60–90 minutes. The 60-90-minute duration was selected to balance realism and organizational feasibility. Tabletop exercises are intended to provide a structured environment in which participants can discuss incident scenarios, evaluate response options, and improve coordination without the resource demands of full-scale operational exercises. A compact format enables participation from technical, managerial, legal, privacy, and communication stakeholders while still supporting scenario escalation, decision-making under uncertainty, and post-exercise reflection [13].

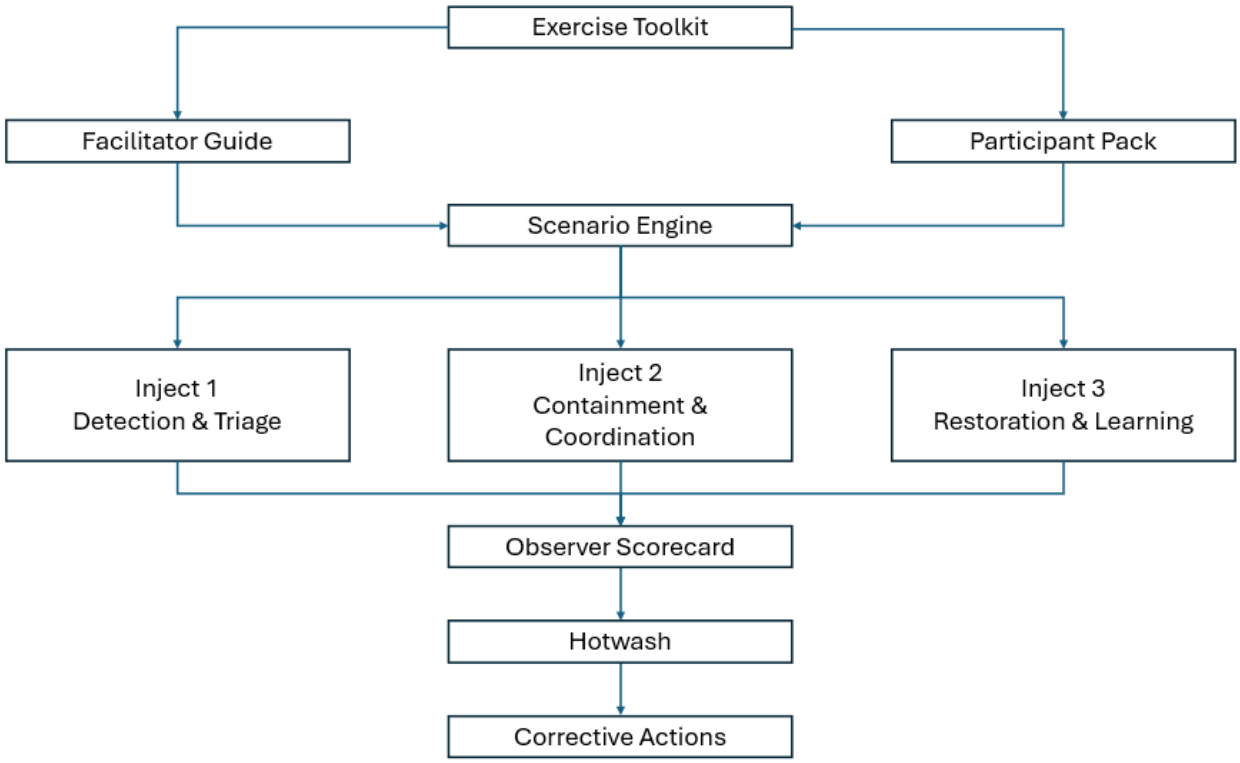


Figure 1. Overall architecture of the Cross-Border Incident Response Toolkit and its learning cycle

The exercise employs six recommended operational roles representing the primary stakeholders typically involved in cross-border digital health incident response. These roles ensure that technical, organizational, legal, and communication perspectives are represented during decision-making. The recommended roles are: (1) Incident Lead – responsible for overall coordination, prioritization, and escalation decisions. (2) Technical Lead – responsible for technical analysis, containment options, dependency assessment, and restoration planning. (3) Gateway/NCP Owner – responsible for interoperability services and cross-border exchange operations. (4) Privacy Officer / DPO – responsible for privacy impact assessment, notification obligations, and regulatory considerations. (5) Communications Lead – responsible for internal and external communications. (6) Vendor Liaison – responsible for interactions with suppliers, evidence requests, support access, and accountability discussions. The role structure was intentionally designed to expose the distributed nature of incident-response responsibilities within interconnected healthcare ecosystems. Rather than focusing solely on technical remediation, participants must coordinate decisions across operational, legal, clinical, and supplier domains. The Facilitator Guide provides exercise-control information, timing guidance, escalation prompts, expected outputs, evaluation criteria, and discussion questions for each stage of the exercise. Its purpose is to ensure consistency across repeated exercise sessions while maintaining realism and controlled progression of the scenario. The Participant Pack provides only the operational information necessary for decision-making. This information asymmetry preserves uncertainty and creates realistic pressure, similar to that of an actual cybersecurity incident. Participants are expected to make decisions without complete information and to identify assumptions and evidence gaps explicitly. Together, the two documents create a reusable training asset that supports both facilitated delivery and participant engagement. The scenario is organized as a progressive sequence of three injects that gradually increase operational complexity and decision pressure. Each inject is designed to emphasize a different phase of incident response while preserving continuity throughout the scenario. Inject 1 -

Detection and Triage: The exercise begins with unusual privileged login activity detected on integration servers at Hospital A, while clinicians at Hospital B report failures in cross-border patient summary requests. No confirmed evidence of data exfiltration is available at this stage. Participants must determine an initial severity assessment, identify immediate evidence requirements, and decide whether service operations continue normally or under restrictions. Inject 2 - Containment and Cross-Border Coordination: The second inject introduces failed gateway transactions, a suspicious supplier support session, and external media interest regarding possible exposure of patient information. Participants must evaluate containment thresholds, establish communication channels across organizations and countries, determine supplier engagement responsibilities, and balance operational continuity against privacy and security concerns. Inject 3 - Restoration and Learning: The final inject introduces forensic findings indicating access to integration logs and configuration data through a supplier account. Although no patient data exfiltration has been confirmed, service restoration is technically possible under enhanced controls. Participants must define restoration criteria, determine notification requirements, document residual risks, and identify corrective actions.

Table 1. Primary Focus Areas Of The Exercise Inject

Inject	Focus Area	Primary Challenge
Inject 1	Detection and Triage	Incident ownership and evidence preservation
Inject 2	Containment and Coordination	Cross-border communication and supplier involvement
Inject 3	Restoration and Recovery	Notifications, residual risk, and service restoration

The exercise is designed to evaluate participants' ability to make defensible decisions under uncertainty rather than to achieve technically perfect outcomes. These objectives directly reflect common operational challenges encountered in interoperable healthcare ecosystems where multiple organizations, jurisdictions, suppliers, and infrastructures participate in service delivery.

The assessment dimensions were derived from two sources: (1) recurring incident-response challenges identified in healthcare cybersecurity literature and (2) the design objectives of the proposed artifact. Prior research highlights the importance of clearly defined responsibilities, effective communication and coordination, timely decision-making, and evidence management during cyber incidents [4], [13], [15]. The proposed toolkit extends these considerations to cross-border digital health environments by explicitly incorporating interoperability dependencies and multi-stakeholder coordination requirements associated with EHDS-enabled health information exchange [3], [6]. The toolkit incorporates a lightweight evaluation mechanism consisting of an observer scorecard and a structured hotwash process. The scorecard focuses on observable behaviors rather than technical correctness. Five dimensions are assessed on a three-point scale ranging from 0 (missing) to 1 (partial) to 2 (clear, timely, and defensible).

The assessed dimensions are:

- Role clarity.
- Cross-border coordination and escalation.
- Containment decision quality under uncertainty.
- Evidence handling and logging priorities.
- Communications, privacy, and regulatory judgment.

Following the exercise, participants conduct a structured hotwash discussion designed to translate observations into concrete improvement actions. The hotwash identifies: (1) Decision bottlenecks, (2) Dependency-related uncertainties, (3) Communication shortcomings, (4) Evidence collection challenges, and (5) Opportunities for risk reduction. The required outputs include the top corrective actions, assigned owners, implementation target

dates, process improvements, technical improvements, and recommendations for subsequent exercises or validation activities. Consequently, the toolkit functions not only as a preparedness assessment instrument but also as a practical organizational improvement mechanism.

V. DESIGN RATIONALE AND ANTICIPATED OUTCOMES

1. Operational Dependency Areas Revealed by the Exercise

The development and scenario analysis process identified four dependency areas that consistently influence cross-border incident response activities in digital health environments: (1) Identity and access management services; (2) Gateway and National Contact Point (NCP) operations; (3) Vendor and third-party access; and (4) Cross-organizational data flows and information exchange mechanisms. These dependencies frequently act as operational bottlenecks during cybersecurity incidents. Although they often remain largely invisible during normal service operation, they become critical when organizations must make rapid decisions under conditions of uncertainty. The exercise repeatedly exposes situations in which participants must determine incident ownership, identify evidence sources, coordinate stakeholders, and balance operational continuity against security and privacy requirements. Based on the design analysis and literature review, these dependency areas are expected to influence incident scope estimation, accountability, evidence collection priorities, containment decisions, communication strategies, and service restoration planning. Consequently, one of the most significant outcomes of the toolkit design is its ability to make interoperability-related dependencies visible and discussable within a structured exercise environment.

2. Evaluation and Learning Benefits

The study resulted in a lightweight evaluation approach integrated into the exercise process. Rather than measuring technical accuracy, the observer scorecard focuses on observable behaviors associated with effective incident response, including role clarity, coordination, communication, evidence handling, and decision-making under uncertainty. The exercise design is intended to encourage participants to justify their decisions, document their assumptions, and communicate uncertainties rather than waiting for complete information. This reflects the realities of cybersecurity incident management, where critical decisions frequently must be made before definitive evidence is available. The separation between facilitator and participant materials is expected to support consistent exercise delivery while maintaining information asymmetry and realistic decision pressure. The three-inject scenario structure progressively increases operational complexity by moving participants from incident detection and triage to containment, cross-border coordination, restoration, and recovery activities. In particular, the inclusion of a supplier-related inject is intended to stimulate discussion regarding accountability, evidence ownership, access rights, and communication responsibilities, thereby highlighting the importance of third-party relationships in interoperable healthcare ecosystems. Consequently, the evaluation approach is expected to provide organizations with a practical mechanism to identify coordination challenges, communication gaps, and decision bottlenecks, while supporting preparedness assessment and organizational learning.

3. Continuous Improvement Potential

Although formal pilot validation has not yet been completed, the current results indicate that the toolkit provides practical value as a preparedness and organizational learning mechanism. The exercise concludes with a structured hotwash process that requires participants to identify corrective actions, assign responsible owners, and define target implementation dates. Typical outputs include: Corrective actions; Responsible owners; Target implementation dates; Process improvement proposals; Technical control enhancements; and Recommendations for future exercises or validation activities. By linking observations directly to improvement activities, the toolkit

extends beyond a discussion-based exercise and provides a practical foundation for continuous enhancement of cyber resilience capabilities. The approach, therefore, supports both preparedness assessment and organizational learning within increasingly interoperable digital health environments.

VI. DISCUSSION

1. Implications for Cross-Border Incident Preparedness

The development of the toolkit highlights a broader challenge facing interoperable healthcare ecosystems: incident response preparedness increasingly depends on coordination across organizational, technical, legal, and national boundaries rather than on the capabilities of individual organizations alone. As healthcare services become more interconnected through initiatives such as the European Health Data Space (EHDS), cybersecurity incidents are likely to involve multiple stakeholders with different responsibilities, priorities, and decision-making processes. From a practical perspective, the proposed toolkit provides a structured environment for examining how these stakeholders interact during a disruptive event. Instead of emphasizing technical problem-solving alone, the exercise encourages participants to consider governance, accountability, communication, privacy obligations, and operational continuity simultaneously. This reflects the multidimensional nature of incident response in contemporary digital health environments. A further implication is that preparedness should be viewed as an organizational capability rather than merely a technical capability. Effective response depends not only on detecting and mitigating cyber threats but also on maintaining shared situational awareness, coordinating decisions, and communicating effectively when information is incomplete or evolving.

2. Contribution of the Research

The primary contribution of this work is the creation of a reusable tabletop exercise artifact specifically designed for cross-border digital health incident response. Existing tabletop exercises frequently address general cybersecurity scenarios or organization-specific incidents, whereas the proposed toolkit explicitly incorporates interoperability dependencies, supplier interactions, and cross-border coordination requirements.

From a Design Science Research perspective, the study demonstrates how interoperability-related cybersecurity challenges can be translated into a practical artifact that supports preparedness activities. The contribution lies less in developing new cybersecurity theory and more in providing a structured mechanism that organizations can use to rehearse and evaluate incident-response practices within complex healthcare ecosystems.

3. Limitations and Future Research

Several limitations should be acknowledged. First, this study reports the design and initial development of the toolkit rather than validated operational outcomes. Its effectiveness has not yet been assessed through pilot exercises involving representative healthcare stakeholders, and the findings remain preliminary. Second, the current scenario represents only one class of cybersecurity incident. Although the scenario was selected to reflect realistic interoperability-related challenges, other relevant threats may require different exercise designs and evaluation criteria. Examples include ransomware incidents, insider threats, cloud service failures, and identity federation disruptions. Third, the observer scorecard was intentionally designed as a lightweight evaluation instrument. While simplicity supports practical adoption, additional research is needed to assess the reliability and repeatability of the evaluation approach across different organizational contexts. Future work should focus on empirically validating the toolkit through pilot exercises with healthcare providers, interoperability service operators, and other stakeholders involved in cross-border health information exchange. Such pilots would enable assessment of usability, realism, participant engagement, and perceived organizational value. A further area for investigation concerns longitudinal use of the toolkit. Repeated exercise cycles could be used to examine

whether structured tabletop exercises contribute to measurable improvements in coordination, communication, decision-making, and overall cyber resilience over time. In addition, future studies could explore integrating tabletop exercises with technical validation activities and broader cyber-resilience assessment frameworks.

VII. CONCLUSION

This work-in-progress paper presents the design and implementation status of a reusable tabletop exercise toolkit for cross-border digital health incident response. The current contribution is a practical foundation rather than a finished solution. It demonstrates that a scenario-specific tabletop toolkit can be created for cross-border digital health incident response and can structure discussion around the dependencies that matter most. Although formal validation remains future work, the current results indicate that the toolkit can support preparedness by making hidden dependencies visible and improving the structure of incident response discussions in interoperable healthcare environments.

ACKNOWLEDGEMENTS

The authors would like to thank the XiA project (funded by the European Union under Grant Agreement No. 101187650) for providing an inspiring and timely research topic, as well as for the valuable guidance and support throughout the research process. The constructive feedback helped shape the model's development and improve its practical relevance.

AI DECLARATION

Artificial intelligence tools, including Copilot, ChatGPT, and Claude, were used to support language editing, structure checking, and clarity improvements during the preparation of this paper. The authors made all ideas, analyses, conclusions, and final editorial decisions.

REFERENCES

[1] K. Kioskli, S. Papastergiou, T. Fotis, S. Silvestri, and H. Mouratidis, 'A Self-Organized Swarm Intelligence Solution for Healthcare ICT Security', Hum. Factors Cybersecurity, 2024, doi: 10.54941/ahfe1004780.

[2] M. Nalin et al., 'The European cross-border health data exchange roadmap: Case study in the Italian setting', J. Biomed. Inform., vol. 94, p. 103183, Jun. 2019, doi: 10.1016/j.jbi.2019.103183.

[3] C. Luidold and C. Jungbauer, 'Cybersecurity policy framework requirements for the establishment of highly interoperable and interconnected health data spaces', Front. Med., vol. 11, May 2024, doi: 10.3389/fmed. 2024.1379852.

[4] K. Nakayama and K. Watanabe, 'Incident response exercises and methodologies to guide best practices for incident response in healthcare institutions', Emerg. Technol. Healthc. Med., 2023, doi: 10.54941/ahfe1004378.

[5] K. Fišter and H. Belani, 'Cybersecurity preparedness and resilience in health care: a narrative review', Int. J. Health Gov., vol. 31, no. 1, pp. 38–49, Nov. 2025, doi: 10.1108/IJHG-07-2025-0108.

[6] C. Stellmach, M. R. Muzoora, and S. Thun, 'Digitalization of Health Data: Interoperability of the Proposed European Health Data Space', in Digital Professionalism in Health and Care: Developing the Workforce, Building the Future, IOS Press, 2022, pp. 132–136. doi: 10.3233/SHTI220922.

[7] EHDS - Art. 23 | StreamLex'. Accessed: Jul. 03, 2026. [Online]. Available: https://streamlex.eu/articles/ehds-en-art-23/

[8] J. Rajamäki and J. Virtanen, 'Leveraging European Health Data Space (EHDS) for Safer and Smarter European Travel', Int. Conf. Tour. Res., vol. 9, no. 1, pp. 395–404, Apr. 2026, doi: 10.34190/ictr.9.1.4490.

[9] Bincoletto, 'Data protection issues in cross-border interoperability of Electronic Health Record systems within the European Union', Data Policy, vol. 2, p. e3, Jan. 2020, doi: 10.1017/dap. 2020.2.

[10] Home - MyHealth@Eu NCPeH API v9.1.1-ci. Accessed: Jul. 03, 2026. [Online]. Available: https://build-fhir.ehdsi.eu/ncp-api/

[11] 2024-12-16_COCIR.EHDS_.Cybersecurity.Interoperability.FINAL_-1.pdf'. Accessed: Jul. 03, 2026. [Online]. Available: https://www.cocir.org/wp-content/uploads/2024-12-16_COCIR.EHDS_.Cybersecurity.Interoperability.FINAL_-1.pdf

[12] Build a Connected Healthcare System. Accessed: Jul. 03, 2026. [Online]. Available: https://www.oracle.com/health/interoperability-healthcare/

[13] G. N. Angafor, I. Yevseyeva, and L. Maglaras, 'Scenario-based incident response training: lessons learned from conducting an experiential learning virtual incident response tabletop exercise', Inf. Comput. Secur., vol. 31, no. 4, pp. 404–426, Mar. 2023, doi: 10.1108/ICS-05-2022-0085.

[14] A. R. Hevner, S. T. March, J. Park, and S. Ram, 'Design Science in Information Systems Research1', Manag. Inf. Syst. Q., vol. 28, no. 1, pp. 75–106, Mar. 2004, doi: 10.2307/25148625.

[15] M. S. Jalali, B. Russell, S. Razak, and W. J. Gordon, 'EARS to cyber incidents in health care', J. Am. Med. Inform. Assoc., vol. 26, no. 1, pp. 81–90, Jan. 2019, doi: 10.1093/jamia/ocy148.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Multimodal Analysis Approach for Automated Classification of Content Rating Symbols in Video Content

Barış Şenel
Computer Engineering Department
Technology Faculty, Gazi University
Ankara, Türkiye
senel18baris@gmail.com

Murat Dörterler
Computer Engineering Department
Technology Faculty, Gazi University
Ankara, Türkiye
dorterler@gazi.edu.tr

Adem Tekerek
Computer Engineering Department
Technology Faculty, Gazi University
Ankara, Türkiye
atekerek@gazi.edu.tr

Abstract
The rapid growth of digital video platforms has heightened the need for scalable, consistent content rating mechanisms. Manual rating processes are often costly, time-consuming, and subjective, especially when video content includes complex visual, auditory, and textual cues. This paper presents an artificial intelligence-supported multimodal analysis approach for the automated classification of content rating symbols. The proposed system analyzes video content across multiple modalities and produces multi-label decisions on violence, horror, sexuality, and bad behavior, along with an age-rating recommendation. The system was evaluated on a curated dataset of 452 videos representing different resolutions, lighting conditions, and content categories. Experimental results indicate that the proposed approach is promising as a human-in-the-loop decision-support tool for content rating, while further optimization is required for fully autonomous deployment.

Keywords: Content Rating Symbols Classification, Multimodal Analysis, Multi-label Classification, Human-in-the-loop Decision Support

I. INTRODUCTION

Digital media distribution has changed the scale of video production and consumption. Traditional broadcast content can be accessed through scheduled programming and a limited number of channels. However, streaming services and social platforms generate large amounts of video content every day. This volume makes manual inspection costly, slow, and difficult to apply consistently. Human reviewers may also be affected by fatigue, subjectivity, and missed violations. Age classification and content warning decisions cannot rely on a single cue. Violence may appear through objects, actions, dialogue, tone, or scene context. Offensive behavior may be verbal rather than visual. A risky object may also be harmless in a specific scene. For this reason, content rating requires evidence from multiple modalities. Visual, auditory, and textual information should be interpreted together. In this study, content rating symbols refer to warning labels for violence, horror, sexuality, and inappropriate behavior. Age rating recommendations also support these labels. The objective of this study is to develop and evaluate a prototype multimodal content analysis system. The system supports semi-automated and human-in-the-loop smart labeling of video content. It aims to assist the automated classification of content rating symbols. For this purpose, it integrates timestamped evidence from different modalities into a structured decision-making process. The main contributions are as follows: (i) an end-to-end multimodal pipeline for extracting visual, object-level, and speech-based evidence from video content; (ii) a decision layer for multi-label classification of content rating symbols; (iii) an operational reliability mechanism for structured decision generation during batch analysis; and (iv) a human-in-the-loop review interface for evidence-based validation and reviewer feedback.

II. RELATED WORK

Early content moderation systems mainly relied on metadata. These systems used titles, descriptions, tags, subtitles, and user-provided labels. Text-based classifiers such as TF-IDF and Naive Bayes have also been used for categorization and filtering tasks [1], [2]. These methods are computationally efficient. However, they depend heavily on the quality of textual information. Metadata may be missing, misleading, incomplete, or intentionally manipulated. For content rating symbols classification, this is a major limitation. Rating decisions should reflect the video's actual visual and auditory content, not just its metadata. Computer vision methods moved content analysis from metadata to frame-level evidence. Early approaches used handcrafted features such as HOG and SIFT. Later studies adopted convolutional neural networks and real-time object detection models [3-6]. Object detectors can identify risk-related visual cues such as weapons, alcohol, or other explicit objects [7]. However, object detection alone is not sufficient for content rating. The same object may have different meanings in different scenes. For example, a knife in a cooking scene and a knife in an assault scene may look similar. Their rating implications, however, are different. This shows the need for contextual interpretation. Audio analysis is also important in video content rating. Harmful or age-inappropriate content may appear through speech, threats, profanity, screams, or emotional tone. Early automatic speech recognition systems were based on statistical models such as hidden Markov models [8]. More recent transformer-based approaches provide stronger transcription performance, especially under noisy and multilingual conditions [9]. Still, speech evidence should not be interpreted in isolation. A loud voice, tense tone, or sports commentary may be incorrectly associated with violence or fear if visual context is ignored [10]. Vision-language models help connect visual perception with semantic interpretation. They generate natural-language descriptions of scenes and provide higher-level context beyond object labels [11], [12]. These descriptions can support content rating by explaining what is happening in a scene. However, scene descriptions, detected objects, and speech transcripts still need to be combined into a final decision. Large language models can support this stage by synthesizing chronological multimodal evidence and producing structured outputs [13]. At the same time, hallucination and prompt sensitivity remain important risks. These risks are especially critical in safety-related systems, where false positives and false negatives may affect rating decisions [14], [15]. Overall, previous studies show that text, vision, and audio-based methods each provide useful but incomplete evidence for content moderation. Single-modality systems may miss contextual cues or produce misleading predictions. This limitation is especially important for classifying content rating symbols. A video may require multiple warning labels, and each label may depend on evidence from different modalities. Therefore, this study focuses on integrating multimodal evidence and providing human-reviewable decision support for multi-label content rating.

III. MATERIALS AND METHODS

1. Dataset and preprocessing

The experimental dataset was designed around four main risk categories: violence, horror, sexuality, and bad behavior. Since a complete, ready-to-use dataset covering all target categories was not available, videos were collected and curated from multiple publicly accessible sources. The final dataset contains 452 video samples. The videos are available in different resolutions, including 720p, 1080p, and 4K. They also vary in lighting conditions, visual quality, and cultural context.

Each video was manually labeled using a multi-label annotation scheme. This allowed a single video to contain multiple risk categories. In addition to the four risk categories, a general content class was included as a non-risk reference category. Table 1 presents the distribution of the dataset across the target categories.

Table 1. Dataset distribution across risk categories

Risk category	Number of videos	Label type
Violence	100	Multi-label
Horror	100	Multi-label
Sexuality	52	Multi-label
Bad behavior	100	Multi-label
General	100	Single label
Total	452	-

Before analysis, each video was decomposed into visual and audio components. Representative frames were extracted from the video stream, and the audio track was converted into a separate waveform file. Frames were sampled at 1 fps to reduce redundant computation and to limit the amount of visual evidence passed to the reasoning stage. This sampling strategy improves processing efficiency. However, it may miss short-duration visual events. The audio stream was exported as a 16 kHz WAV file to support speech transcription.

2. Multimodal analysis pipeline

The proposed architecture combines visual, auditory, and textual evidence within a unified analysis pipeline. First, sampled video frames are analyzed to detect risk-related visual cues and objects. Second, the same frames are used to generate scene-level descriptions. These descriptions provide contextual information about actions, relationships, and visual situations that object detection alone may not capture. Third, the audio stream is transcribed into timestamped text. The outputs of these modules are then merged into a chronological evidence timeline. Figure 1 shows the proposed multimodal video analysis and decision pipeline.

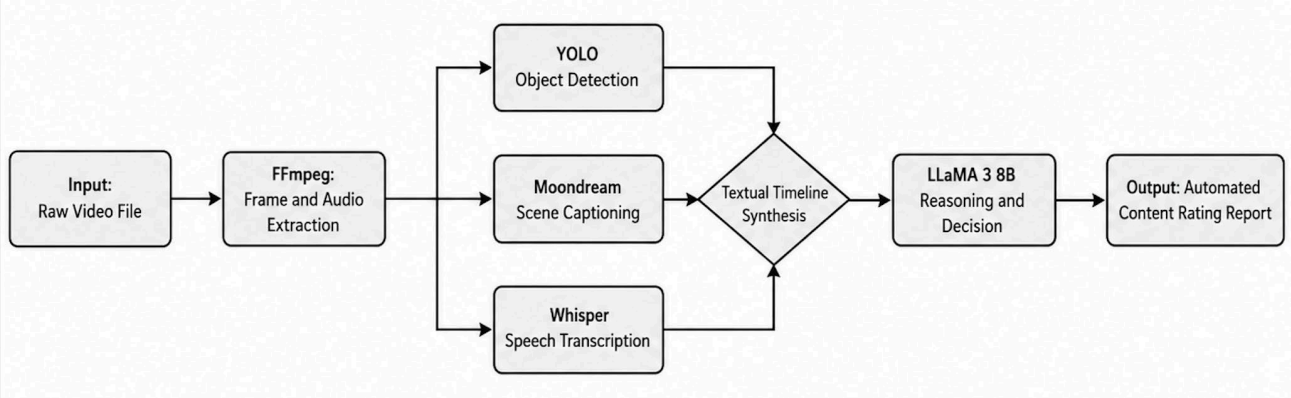


Figure 1. Proposed multimodal video analysis and decision pipeline.

The chronological evidence timeline is passed to the decision-making component. This component follows a rule-based rating prompt and assigns independent labels for violence, horror, sexuality, and bad behavior. It also provides an age-rating recommendation and a brief explanation. The LLaMA 3 model was not fine-tuned in this study. Instead, its decision behavior was adapted through prompt engineering. The decision prompt simulated a rule-based content-rating process and included negative constraints to reduce prompt-induced

hallucinations and overly sensitive labeling. This design supports interpretable decision-making. Rather than producing only a final class label, the system provides evidence-based outputs for human evaluators to review.

3. Robust decision generation and error handling

Structured output generation is important for batch-based content analysis. During long-running experiments, the language model may occasionally return outputs that do not fully conform to the requested JSON format. Such formatting errors may disrupt the processing pipeline if not handled properly. To address this problem, the system includes a Python-based recovery mechanism. This mechanism attempts to extract the most probable JSON block using regular expressions. It also repairs missing brackets and simple string-formatting errors when possible. If recovery is not possible, the system returns a safe fallback structure. This mechanism does not directly improve classification accuracy. However, it improves the operational reliability of the batch-processing pipeline. The conducted experiment allowed the complete dataset to be processed for 4.5 hours without a system crash.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

1. Experimental setup

Experiments were conducted in Google Colab using an NVIDIA A100 Tensor Core GPU. This environment provided sufficient computational capacity for running the multimodal analysis pipeline. The implementation was developed in Python and executed using PyTorch and Hugging Face Transformers. The main reasoning and vision-language components were loaded without 4-bit or 8-bit quantization. This configuration was preferred to preserve the quality of semantic interpretation during multimodal evidence analysis. The models were executed with bfloat16 precision to balance computational efficiency and inference quality. The experimental pipeline included frame extraction, object detection, visual captioning, speech transcription, multimodal evidence synthesis, and structured decision generation. Each video was processed through the same pipeline to ensure consistency across the dataset. The generated outputs were stored in a structured format for later evaluation and error analysis.

2. Multi-label classification performance

Since a video may contain multiple harmful-content categories, the evaluation was conducted as a multi-label classification task. Precision, recall, and F1-score were computed separately for each risk category. Micro, macro, weighted, and sample averages were also calculated to summarise overall performance (Table 2). The general class was used as a non-risk reference category during dataset construction. It was not included as a target class in the multi-label risk classification metrics reported in Table 2.

Table 2. Multi-label classification results on the test dataset

Class	Precision	Recall	F1-score	Support
Violence	0.40	0.46	0.43	100
Horror	0.69	0.46	0.55	100
Sexuality	0.43	0.58	0.50	52
Bad behavior	0.29	0.93	0.44	100
Micro avg.	0.38	0.61	0.47	352
Macro avg.	0.45	0.61	0.48	352

Weighted avg.	0.46	0.61	0.48	352
Samples avg.	0.30	0.48	0.35	352

The system achieved micro F1 Scores of 0.47 and macro F1 Scores of 0.48. These results indicate moderate preliminary classification performance. They also suggest that further optimization is needed before the system can be used as a fully automated rating mechanism. The horror class achieved the highest precision, at 0.69. This shows that the model was more selective when assigning the horror label. However, its recall was 0.46, which indicates that some horror-related samples were missed. The sexuality class achieved the highest F1-score after horror, at 0.50. The violence class showed lower and more balanced precision and recall values. The bad behavior class produced a different pattern. It achieved a high recall of 0.93 but a low precision of 0.29. This means the system detected most bad-behavior samples, but also assigned this label to many non-relevant samples. This result suggests a risk-averse tendency in labeling. Such behavior may be acceptable in a reviewer-assistance scenario, where a human reviewer later checks suspicious content. However, it would not be sufficient for fully autonomous content rating. Figure 2 presents the independent confusion matrices for the multi-label risk categories. These matrices provide a class-level view of the model's error patterns and help explain the precision-recall differences reported in Table 2.

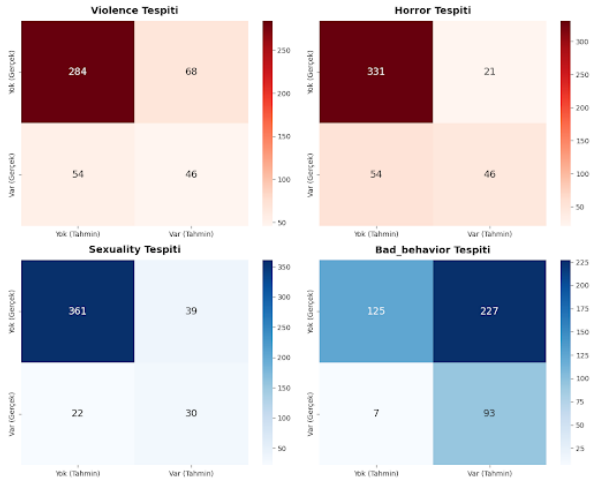


Figure 2. Independent confusion matrices for multi-label risk classes.

3. Error analysis and prompt optimization

Error analysis showed that one of the main sources of false positives was prompt-induced over-sensitivity. In the initial experiments, instructions related to alcohol, smoking, violence, or similar risk indicators caused the model to over-interpret some ordinary visual cues. For example, non-risk objects such as water bottles or background textures were sometimes treated as risky evidence. Similar errors were observed in sports scenes, where physical contact was occasionally classified as violence. In some cases, tense audio tone was also interpreted as horror, even when there was no supporting visual evidence. To reduce these errors, the final prompt was revised with negative constraints. These constraints required explicit and modality-supported evidence before assigning a risk label. The revised prompt encouraged the model to avoid assigning a category based on weak, ambiguous, or isolated cues. This modification shifted the decision process from aggressive labeling to

more conservative, evidence-based reasoning. As a result, the system became better aligned with a human-reviewable content rating workflow.

4. Human-in-the-loop interface

Content rating is a high-stakes task. Therefore, final decisions should remain subject to review by human experts. For this reason, the proposed system includes a web-based human-in-the-loop interface. Figure 3 illustrates the smart labeling interface that presents the recommended age rating, predicted content-rating symbols, explanations, and timestamped evidence to reviewers. The interface presents the recommended age rating, predicted content rating symbols, a general explanation, and timestamped evidence. This allows reviewers to inspect suspicious moments directly. As a result, they do not need to watch the entire video in every case. The interface also supports reviewer feedback. Reviewers can confirm correct predictions or mark false positives. This feedback can be used in future work to improve prompt design, reduce over-sensitive predictions, and support model refinement. In this way, the interface improves transparency and makes the system better suited to reviewer-assistance scenarios.

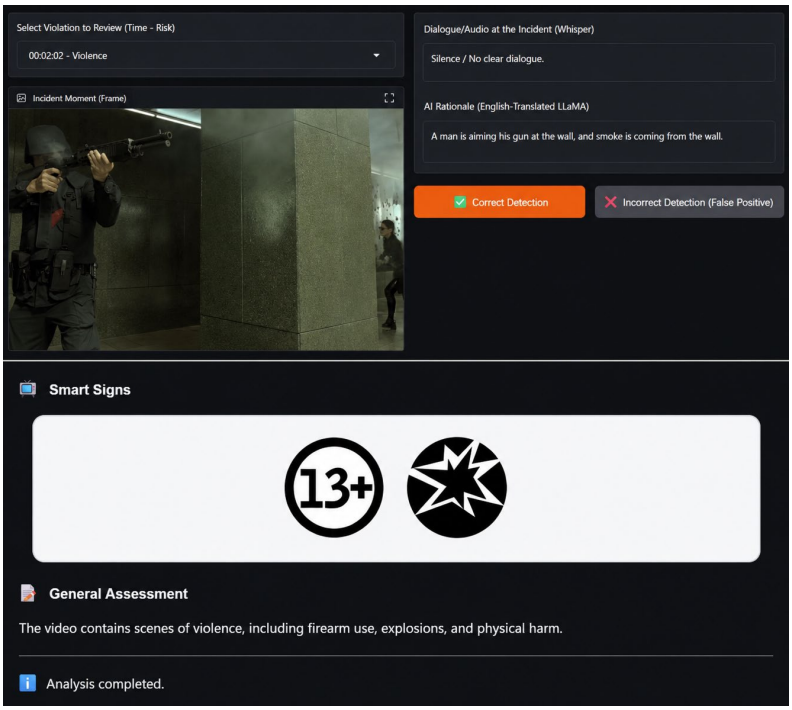


Figure 3. Smart labeling interface showing the predicted age rating and risk indicators.

V. CONCLUSION AND FUTURE WORK

This paper presented an artificial intelligence-supported multimodal content analysis system for the automated classification of content rating symbols in video content. The proposed prototype integrates video preprocessing, object detection, visual scene description, speech transcription, and reasoning-based decision-making into an end-to-end pipeline. The system transforms heterogeneous video evidence into a chronological timeline. It then generates multi-label risk decisions, an age-rating recommendation, and an explanatory rationale. Evaluation of 452 videos showed that the proposed architecture supports offline batch analysis in

reviewer-assistance scenarios. The results also demonstrated the potential of combining visual, auditory, and textual evidence to detect harmful content categories. However, the obtained performance indicates that the system is not yet suitable for fully autonomous deployment. In particular, high recall in sensitive categories may lead to more false positives. This was especially observed when the prompt overemphasized specific risk indicators. Negative prompt constraints and structured-output recovery improved operational reliability. The human-in-the-loop interface also supported transparent validation by presenting timestamped evidence to reviewers. Future work should focus on domain-specific fine-tuning using expert annotations and regulatory decisions. Larger multilingual and culturally diverse datasets are also needed. Real-time stream processing can be implemented using queue-based architectures and optimized inference pipelines. Model compression techniques, such as quantization and accelerated inference, may further improve the feasibility of deployment. Finally, reviewer feedback collected through the interface can be integrated into a continual-learning workflow. This would allow the system to improve over time under expert supervision.

ACKNOWLEDGEMENT

The authors thank Gazi University, Faculty of Technology, Department of Computer Engineering, for the academic environment in which the graduation project underlying this study was developed.

REFERENCES

[1] K. Filippova and K. B. Hall, “Improved video categorization from text metadata and user comments,” in Proceedings of SIGIR, 2011, pp. 835-842
[2] V. Dogra et al., “A complete process of text classification system using state-of-the-art NLP models,” Computational Intelligence and Neuroscience, 2022.
[3] M. Potthast et al., "Clickbait detection," in Proceedings of the European Conference on Information Retrieval, 2016, pp. 810-817.
[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886-893.
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.
[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779-788.
[7] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," Neurocomputing, vol. 275, pp. 66-72, 2018.
[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.
[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2022.
[10] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," Pattern Recognition Letters, vol. 65, pp. 22-28, 2015.
[11] Vikhyat, "Moondream: A tiny vision-language model," GitHub repository, 2024. [Online]. Available: https://github.com/vikhyat/moondream.
[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156-3164.
[13] Meta AI, "Introducing LLaMA 3: The next generation of our open source large language model," Meta Research, 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/.
[14] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55,

no. 12, pp. 1-38, 2023.
[15] N. McKenna, T. Li, L. Cheng, and M. Hosseini, "Sources of hallucination by large language models on inference tasks," in Advances in Neural Information Processing Systems, 2023.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

AI-Supported Emission Factor-Based Approach for Carbon Footprint Analysis and Intersection Prioritization in Sustainable Cities

Elif YILDIZ
Department of Digital Forensics
İstanbul University-Cerrahpaşa
İstanbul, Türkiye
elif.yildiz@iuc.edu.tr
0009-0009-2262-0167

Deniz BALTA
Department of Software Engineering
Sakarya University
Sakarya, Türkiye
ddural@sakarya.edu.tr
0000-0001-9104-1868

Musa BALTA
Department of Cyber Security Engineering
Sakarya University
Sakarya, Türkiye
mbalta@sakarya.edu.tr
0000-0002-8711-6625

Abstract
A growing population and rapid urbanization are increasing traffic congestion in urban transportation systems, leading to higher fuel consumption, a larger carbon footprint, and increased air pollution; this situation heightens the need for emissions-focused decision support systems for sustainable traffic management. In this study, an emissions-focused decision support approach was developed using real-time traffic data collected from seven critical arterial intersections in the Adapazarı district of Sakarya Province, in collaboration with the Sakarya Metropolitan Municipality. Using the traffic data, CO_2, NO_x, and exhaust-derived $PM_{2.5}$ emissions were calculated with the COPERT/EMEP-EEA Tier 2 methodology; subsequently, a normalized multi-pollutant emission priority score was developed to evaluate pollutants collectively. Using the resulting emission scores, critical intersection-time combinations were identified, and emission levels were classified using the Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost algorithms. Model performance was evaluated using the Leave-One-Group-Out cross-validation method, and the best results were obtained with the XGBoost algorithm. The XGBoost model demonstrated superior performance compared to other methods, with an accuracy of 0.978, an F1 score of 0.972, and an RMSE value of 0.147. The proposed approach integrates intersection traffic data, spatial impact distances, COPERT-based emission estimates, multi-pollutant emission scoring, Top-K prioritization, and machine-learning-based emission classification into a unified decision-support framework for sustainable cities, rather than evaluating traffic density in isolation.

Keywords: AI-Powered Traffic Management, Sustainable Urbanization, Carbon Footprint Reduction, COPERT Tier 2, Decision Support System.

I. INTRODUCTION

Rapid urbanization, population growth, and increasing vehicle ownership have intensified traffic congestion in urban transportation systems, leading to higher fuel consumption, increased greenhouse gas emissions, and deteriorating air quality. Intersections, where stop-and-go traffic frequently occurs, are among the most critical sources of transportation-related emissions and therefore play an important role in sustainable traffic management [1],[2]. Emissions from urban transportation are of environmental importance for greenhouse gases that cause climate change and air pollution that negatively affects human health [3]. In this context, while CO_2 emissions are linked to carbon footprints and climate impacts, pollutants such as NO_x and $PM_{2.5}$ are primarily assessed in terms of urban air quality, respiratory health, and local environmental impacts [4]. The Paris Climate Agreement, signed in 2015, and the EU Green Deal aim to significantly reduce emissions by 2030. However, projections indicate that Turkey's CO_2 emissions in 2030 could be 42% above the target level (approximately 723.97 million metric tons) [5]. It has been argued that even if all cars in Turkey were replaced with electric vehicles, CO_2 emissions from passenger cars would decrease by 48.9%. Yet, the resulting

reduction in total vehicle emissions would be limited to 9% [6]. Current literature on traffic management and intersection analysis mainly focuses on transportation performance indicators such as traffic volume, average speed, and travel time. Emissions modelling studies, by contrast, often calculate pollutant concentrations or use macroscopic traffic simulations, but rarely use emissions data to prioritize intersections for decision support. Few studies combine COPERT-based emission calculations with multi-pollutant prioritization and machine learning-based emission classification. There is a clear need for traffic management systems that prioritize emissions and integrate environmental impacts into decision-making, rather than relying only on traditional methods to reduce traffic volume. Developing such systems is crucial for sustainable urban planning. This study analysed traffic data for seven arterial intersections in the Adapazarı district of Sakarya Province on an intersection-hour basis. Intersection locations were represented using the Manhattan distance to calculate influence distances, yielding more descriptive traffic activity values for each intersection. While Euro 6 vehicle technology was selected and decarbonization of transportation is the goal, this could limit such gains, as the number of kilometers travelled per vehicle is expected to increase [4]. The COPERT/EMEP-EEA Tier 2 approach was chosen because it uses emission factors specific to vehicle class and emission standards, is consistent with European Emission Inventory guidelines, and can provide reliable emission estimates without requiring detailed microscopic driving data. Thanks to these features, the method provides a suitable framework for urban emission analyses and decision-support applications based on real traffic data [7]. The calculated CO_2, NO_x, and $PM_{2.5}$ emission values, which vary in units and magnitudes, were evaluated for decision support, normalized, and used to create an emission priority score. This score ensures that intersection-hour observations are ranked not only by traffic volume but also by emission load, accounting for multiple pollutants simultaneously [8]. Traffic-related emissions are affected not only by traffic volume, but also by temporal variation, intersection location, road segment length, and pollutant type. To address these complexities, this study proposes a decision-support workflow that converts hourly traffic counts to vehicle-kilometer activity, estimates CO2, NOx, and $PM_{2.5}$ emissions, and classifies intersection-hour observations by emission intensity using supervised machine learning models. In this context, the study's key contributions to the literature can be summarized as follows:

- CO_2, NO_x, and $PM_{2.5}$ emissions are estimated using the COPERT/EMEP-EEA Tier 2 methodology based on real-time traffic data.
- A normalized multi-pollutant emission priority score is proposed for jointly evaluating different pollutants.
- A Top-K prioritization strategy is introduced to identify critical intersection-hour combinations according to emission load.
- Multiple supervised machine learning algorithms are compared using Leave-One-Group-Out cross-validation to evaluate their generalization performance.
- An integrated decision-support framework is presented for environmentally sustainable traffic management.

II. RELATED WORK

Modelling road transportation emissions and developing sustainable traffic management strategies have become increasingly important due to rapid urbanization, population growth, and climate change [9]. Existing studies generally fall into two categories: (i) artificial intelligence-based traffic and emission prediction [10],[11], and (ii) emission estimation using macroscopic traffic and emission models [12-15]. Although both research directions have achieved significant progress, they are typically investigated independently. Also, in studies using macroscopic traffic models, emission estimates are made on a very large-scale using data such as the average vehicle speed, intersection length, and vehicle composition -in addition to the number of vehicles- but they provide limited information because they do not delve into detail in more detailed studies [1],[13],[16],[10].

Among emission estimation approaches, COPERT and HBEFA are the most widely adopted models in Europe [10], [17]. COPERT estimates emissions using vehicle-specific emission factors and average traffic conditions, whereas HBEFA additionally requires detailed traffic-state information. These studies show that the COPERT method can provide reliable emission estimates at different scales. However, most of these studies have focused on emission calculations, and the integration of these emissions into intersection prioritization or decision support systems has not been addressed in detail [2],[13],[17],[18]. Furthermore, while HBEFA requires detailed traffic condition information, COPERT Tier 2 can produce reliable results with limited data. For this reason, COPERT Tier 2 was deemed more suitable for this study, which is based on actual traffic counts. Machine learning has also been extensively applied to traffic and emission prediction. Tree-based methods such as Random Forest and XGBoost have shown strong performance in modeling nonlinear traffic relationships, while Support Vector Machines and deep learning approaches have been successfully applied to carbon emission forecasting and environmental monitoring [5], [19-21]. Nevertheless, many deep learning approaches require large-scale datasets and substantial computational resources. Considering the relatively limited size and node-based structure of the dataset used in this study, classical machine learning algorithms provide a more interpretable and computationally efficient alternative. Several studies have emphasized that sustainable transportation planning should consider multiple environmental indicators rather than relying solely on traffic density. Multi-criteria prioritization approaches have been proposed for urban planning and environmental assessment, while inventory-based studies have demonstrated the importance of evaluating pollutants such as CO_2, NO_x, and $PM_{2.5}$ simultaneously [8], [22-24]. Despite these advances, existing studies rarely combine multi-pollutant emission estimation, intersection prioritization, and machine learning within a unified decision-support framework. Although significant progress has been made in the literature on emissions modeling, traffic forecasting, and AI-based traffic management, these approaches are mostly addressed independently of one another. Research on decision support systems that integrate multi-pollutant emission calculations using real-time intersection data, critical intersection-hour prioritization, and machine-learning-based emission classification is quite limited. This study addresses these gaps in the literature by integrating the COPERT Tier 2 emissions model, a multi-pollutant emissions score, the Top-K prioritization approach, and supervised machine learning models into a single analytical framework.

III. METHODOLOGY

1. Dataset Definition

The dataset contains real-time traffic data from seven key intersections in central Sakarya, Türkiye, gathered in partnership with the Sakarya Metropolitan Municipality. The intersections represent main arterial roads with diverse traffic patterns. Hourly traffic counts were collected over seven days (24–30 December 2025). Each entry includes the day, hour, intersection name, vehicle count, and additional variables. After cleaning and removing incomplete data, 504 valid observations remained, each signifying a unique intersection-hour. Spatial coordinates of the intersections were also used in the analysis. The traffic data were reformatted, so each record shows a specific intersection, hour, and day. While the dataset covers a longer period, this study centers on three representative days to illustrate the procedure. To capture different traffic conditions, low, medium, and peak periods were identified using average hourly traffic volumes across all intersections. The representative-hour method enabled comparison of emissions under these altering conditions. Since network-level decisions are required, the average total vehicle load across the seven intersections was calculated for each hour. The intervals 04:00–05:00, 07:00–08:00, and 17:00–18:00 were chosen to represent low, medium, and peak traffic, respectively. The intersections were also ordered sequentially along the corridor, and the distances between consecutive intersections were estimated. While Euclidean distance measures the direct linear separation, Manhattan distance more accurately reflects the grid-like configuration of urban road networks. Therefore, the Manhattan method was selected to calculate the representative influence distance between intersections, as urban road geometry commonly involves both horizontal and vertical movements [25]. The influence distance

for each intersection was defined as half the distance to adjacent intersections, which prevents overlapping road segments during activity calculations.

2. Calculating Activity Value and Determining Intensity Classes

CO_2, NO_x, and $PM_{2.5}$ emissions were estimated using the COPERT/EMEP-EEA Tier 2 methodology, in which vehicle-kilometer activity is used as the activity indicator. Accordingly, the activity value was calculated by multiplying the hourly vehicle count by the representative influence distance of each intersection, as defined in Equation (1).

$$Vehicle - km_{j,t} = N_{j,t} \times d_j \quad (1)$$

Where $Vehicle - km_{j,t}$ represents the vehicle-kilometer, $N_{j,t}$ represents the total number of vehicles observed at the relevant intersection during that time; and d_j represents the influence radius of the relevant intersection.

Traffic conditions were categorized into low, medium, and peak periods using the representative-hour approach described in the previous subsection. Following the COPERT Tier 2 methodology, weighted emission factors were computed according to the assumed vehicle fleet composition (80% passenger cars, 15% light commercial vehicles, and 5% heavy-duty vehicles), all complying with the Euro 6 standard, as shown in Equation (2).

$$EF_i^* = \sum_{c=1}^{m} p_c \times EF_{i,c} \quad (2)$$

Where EF_i^* is the weighted emission factor; p_c, denotes the proportion of vehicle category c; $EF_{i,c}$ is the pollutant-specific emission factor of vehicle category i.

For each intersection-hour observation, the emission value was calculated by multiplying the vehicle-km activity by the weighted emission factor for the relevant pollutant. The emission values for CO_2, NO_x, and $PM_{2.5}$ shown in Table 1 were obtained by multiplying the activity value calculated for each observation by the weighted average emission factor for the relevant pollutant.

Table 1. Results of Emission Values at Intersections

Day	**Time Interval**	**Intersection Name**	**Hourly Vehicle Count**	**CO_2**	**NO_x**	**$PM_{2.5}$**
Sunday	00:00 – 01:00	1. Gate	721	82066,14	18,62	0,18
Sunday	04:00 – 05:00	1. Gate	185	49171,39	11,16	0,11
Sunday	07:00 – 08:00	2. Gate	485	35590,13	8,07	0,08
Monday	17:00 – 18:00	3. Gate	3370	304382,58	69,09	0,69
Monday	04:00 – 05:00	4. Gate	94	13051,28	2,96	0,02
Monday	16:00 – 17:00	Boulevard	4310	580366,72	131,74	1,31
Monday	17:00 – 18:00	Boulevard	4179	562726,81	127,73	1,27
Tuesday	07:00 – 08:00	Hospital	1707	157279,92	35,70	0,35
Tuesday	03:00 – 04:00	Y. Mosque	815	41238,08	9,36	0,09

Since CO_2, NO_x, and exhaust-related $PM_{2.5}$ emissions vary in magnitude and impact, a min-max normalization was applied to the emission values for each pollutant. This made the values comparable and enabled their use together in calculating the emission priority score. Following normalization, the emission values were combined using the weighted-sum method to generate a single emission-priority score.

3. Critical Intersection-Time Prioritization

In this study, a Top-K ranking was applied to identify the most critical intersection-time combinations for observation, with the aim not only of determining emission calculations but also of analyzing intervention priorities based on the observations. In this context, the selections were examined separately with respect to the number of vehicles, vehicle-km activity, and multi-pollutant emission scores. This K value was set at 20 to create a manageable set of candidates representing the critical intersection-time combinations that municipalities could address in the short term. The extent to which different selection strategies covered the relevant emission load was calculated using the capture rate shown in Equation 3:

$$R_i = \frac{\sum_{(j,t)\in\mathcal{K}} E_{i,j,t}}{\sum_{j,t} E_{i,j,t}} \tag{3}$$

Here R_i: i represents the capture rate of the critical intersection-time observations selected for pollutant within the total emissions, $\mathcal{K}$: Represents the set of critical intersection-time observations selected using the Top-K prioritization approach, and K has been set to 20. j, t: Represents a specific combination of intersection and time iteral. $E_{i,j,t}$ represents the emission quantity calculated for pollutant at intersection j during time interval t. $\sum_{(j,t)\in\mathcal{K}} E_{i,j,t}$ represents the total emission quantity for the relevant pollutant from the critical intersection-time observations selected using the Top-K method. The top 21 critical observations belong to the Boulevard Intersection, and the remaining observations are also provided in Table 2 to show the values for all intersections:

Table 2. Emission Scores and Criticality Orders for Intersections

Day	**Time Interval**	**Intersection Name**	**Emission Score**	**Criticality Order**
Monday	16:00 - 17:00	Boulevard	1	1
Tuesday	15:00 - 16:00	1. Gate	0,782098926	22
Tuesday	17:00 -18:00	4. Gate	0,706018464	31
Monday	10:00 - 11:00	3. Gate	0,6157047	62
Monday	15:00 - 16:00	Hospital	0,576477651	69
Monday	15:00 - 16:00	Y. Mosque	0,231558427	234
Monday	08:00 - 09:00	2. Gate	0,193773	269

4. Supervised Learning for Emission Level Classification and Performance Metrics

Traffic data, influence distances, start times, days, intersections, and density classes were used as input variables for the classification models. Vehicle-km activity, CO_2, NO_x, and $PM_{2.5}$ emissions, and the emission score were not included in the model inputs because they were directly correlated with the structure of the reference class. This approach reduced the likelihood of data leakage by eliminating variables that directly determine the models' target class. Therefore, the relevant variables were used only during the label creation phase. The hyperparameters used for the classification algorithms are presented in Table 3. Since the main aim of this study was to evaluate the applicability of the proposed emission-oriented decision-support workflow rather than to optimize a single classifier, predefined algorithm-specific hyperparameter settings were used for all models.

Random Forest was configured with 300 estimators, a minimum leaf size of 2, and balanced class weights; SVM used an RBF kernel with C=1.0, gamma set to scale, balanced class weights, and probability estimation; Gradient Boosting used 200 estimators, a learning rate of 0.05, and a maximum depth of 3; and XGBoost was configured for three-class classification with 300 estimators, a learning rate of 0.05, maximum depth of 4, subsample and colsample values of 0.90, and multiclass log-loss as the evaluation metric. These predefined settings provided a consistent and interpretable basis for comparing model performance, while systematic hyperparameter optimization is left for future work.

Table 3. Hyperparameter Values of Algorithms

Algorithms	**Hyperparameters**
Random Forest	n_estimators=300; min_samples_leaf=2; class_weight=balanced; random_state=42
Support Vector Machine	kernel=RBF; C=1.0; gamma=scale; class_weight=balanced; probability=True; random_state=42
Gradient Boosting	n_estimators=200; learning_rate=0.05; max_depth=3; random_state=42
XGBoost	objective=multi:softprob; num_class=3; n_estimators=300; max_depth=4; learning_rate=0.05; subsample=0.90; colsample_bytree=0.90; eval_metric=mlogloss; tree_method=hist; random_state=42

Leave-One-Group-Out cross-validation was used as an external evaluation protocol to assess the model's performance generalizability across days, with the day variable defined as the group. This approach prevents observations from the same day from appearing in both the training and test sets, thereby avoiding overly optimistic performance estimates that could result from temporal dependence. As shown in Figure 1, the XGBoost algorithm stands out among the models, achieving an accuracy of 0.978, an F1 score of 0.972, and the lowest RMSE of 0.147.

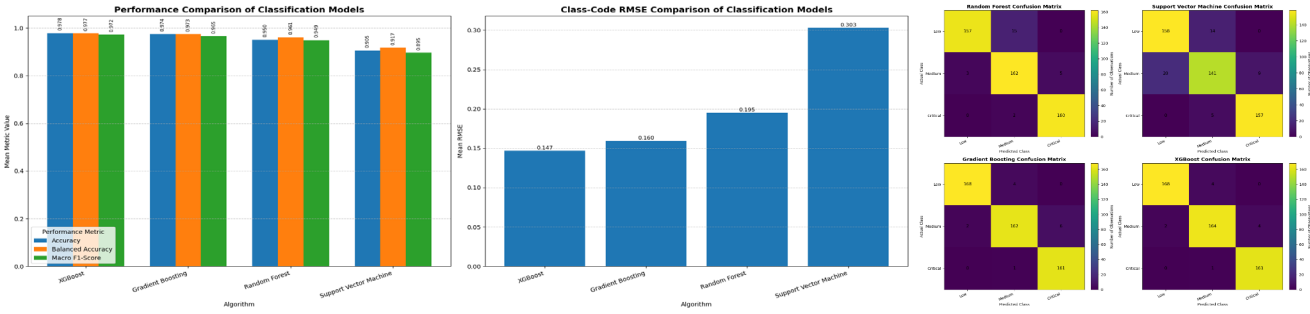


Figure 1. Performance Comparison by Mean Metric Values, Mean RMSE, and Confusion Matrix Results of Algorithms

The confusion matrices in Figure 1 show that the XGBoost and Gradient Boosting models effectively distinguish between low, medium, and critical emission classes, with XGBoost achieving the highest classification performance. Specifically, XGBoost correctly classified 168 of 172 low-emission observations, 164 of 170 medium-emission observations, and 161 of 162 critical-emission observations. Most misclassifications occurred between the low- and medium-emission classes, reflecting the relatively small differences in emission levels between these categories.

IV. RESULTS

In this study, the sample obtained from the dataset was evaluated based on intersection-hour observations spanning three days, using classifications of traffic activity, emission load, and emission levels. In light of the results obtained, as seen in the feature importance evaluations based on different classification algorithms in Figure 2, it is evident that analysis using only traffic data is insufficient; in addition to these data, the influence distance and the spatial impact of traffic mobility on the road network must also be included in the calculation. In other words, the feature importance analysis shows that, in addition to the number of vehicles, the intersection influence radius also plays a significant role in determining emission levels. This finding highlights that assessments based solely on traffic volume may be insufficient in explaining environmental impacts.

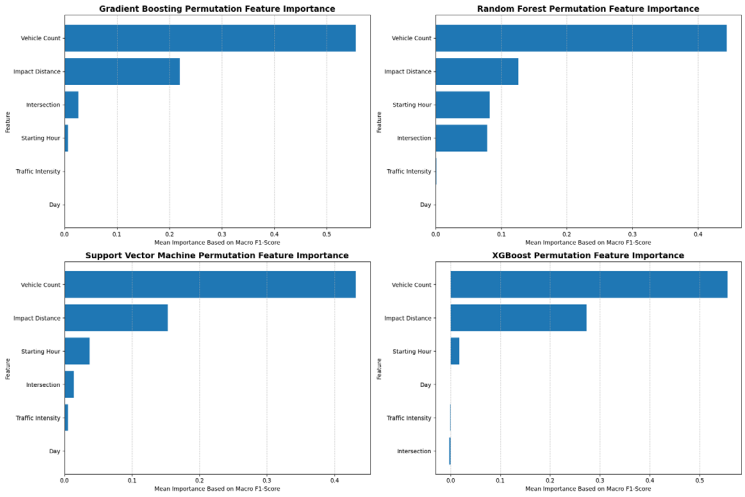


Figure 2. Feature Importance Determined by Models According to the Permutation Method

When all intersections were evaluated, the Boulevard Intersection stood out for both total vehicle count and vehicle-km activity. It was identified as the critical point where the highest total emission load occurs. The Boulevard Intersection's highest emissions across all evaluation criteria indicate that it should be prioritized for traffic management and signal optimization interventions. This approach helps ensure that limited resources are allocated to areas with the greatest environmental impact. When comparing classification models, XGBoost generally demonstrated the highest performance, followed by the Gradient Boosting model. The XGBoost model's superior performance can be attributed to its ability to model nonlinear relationships and learn complex interactions among features more effectively. The nonlinear dependencies among traffic density, intersection location, and time information have contributed to the tree-based boosting approach's superior performance. Furthermore, the findings are consistent with studies in the literature reporting that XGBoost-based models demonstrate high performance in traffic and emissions forecasting [26-28]. However, the proposed study distinguishes itself from existing studies by not only performing emissions forecasting but also incorporating multi-pollutant emissions prioritization into the decision-support process. Overall, the findings indicate that evaluating traffic data alone is insufficient to explain environmental impacts. The combined consideration of traffic density, spatial impact distance, and multi-pollutant emission analyses helps identify more accurate intervention priorities for sustainable traffic management. The study's fundamental contribution to the literature lies in providing an integrated analytical framework that goes beyond merely calculating urban emissions to determine at which intersections and during which time intervals emission interventions should be prioritized. Furthermore, the proposed approach enables municipalities to prioritize signal optimization, traffic routing, and emission reduction strategies by identifying the intersections and time intervals with the highest environmental impact, rather than intervening at all intersections simultaneously.

V. FUTURE WORK

Future studies plan to validate the method on broader road networks and across different cities, integrate it with real-time traffic control systems, and evaluate it comparatively against other emission modeling approaches. This will enable a more comprehensive demonstration of the scalability of the proposed decision-support approach and its performance across various traffic scenarios.

ACKNOWLEDGEMENT

This study was supported by the Scientific Research Projects Coordinator of Sakarya University under grant number 2026-29-86-161 and the Scientific and Technological Research Council of Turkey (TUBITAK) with project number 225E071.

REFERENCES

[1] A. Gedik, O. Uslu, and A. H. Lav, "A prospective study to evaluate CO2 emission mitigation strategies for highway transportation," Environmental Monitoring and Assessment, vol. 194, no. 10, p. 703, Aug. 2022.

[2] N. Abdull, M. Yoneda, and Y. Shimada, "Traffic characteristics and pollutant emission from road transport in urban area," Air Quality Atmosphere & Health, vol. 13, no. 6, pp. 731–738, May 2020.

[3] J. Lu, B. Li, H. Li, and A. Al-Barakani, "Expansion of city scale, traffic modes, traffic congestion, and air pollution," Cities, vol. 108, p. 102974, Oct. 2020.

[4] C. Lu, S. Huang, S. Gao, J. Fu, X. Tian, and S. Lin, "Real-traffic emissions of CO, NOx, CO2, and PM2.5 from vehicles using a portable emission measurement system," Air Quality Atmosphere & Health, vol. 18, no. 8, pp. 2265–2276, Jun. 2025.

[5] M. Kayakuş, M. Terzioğlu, D. Erdoğan, S. A. Zetter, O. Kabas, and G. Moiceanu, "European Union 2030 carbon emission Target: The case of Turkey," Sustainability, vol. 15, no. 17, p. 13025, Aug. 2023.

[6] C. Coskun and Z. Oktay, "Carbon footprint prediction of vehicle usage in Turkey," Greenhouse Gases Science and Technology, vol. 10, no. 4, pp. 736–758, Apr. 2020.

[7] European Environment Agency, EMEP/EEA Air Pollutant Emission Inventory Guidebook 2023—Update 2025, COPERT ver. 5.9, Jun. 2025. [Online]. Available: https://copert.emisia.com/wp-content/uploads/2025/10/1.A.3.b.i-iv-Road-transport-2025.pdf. Accessed: Jul. 1, 2026.

[8] D. Yang, S. Zhang, T. Niu, Y. Wang, H. Xu, and K. M. Zhang, "High-resolution mapping of vehicle emissions of atmospheric pollutants based on large-scale, real-world traffic datasets," Atmospheric Chemistry and Physics, vol. 19, no. 13, pp. 8831–8843, Jul. 2019.

[9] M. A. Yerlikaya, K. Yıldız, B. N. Keskin, and A. Çinar, "Assessing the impact of PM10 and CO2 greenhouse gases from road transportation on global climate change: A mathematical modeling approach," Natural Resources Forum, Jan. 2026.

[10] B. Bhupal, M. Sharma, H. Yadav, V. Sengar, S. Jain, and S. Jain "Ai-driven spatiotemporal urban road transport emission mapping using YOLO and COPERT model," Transportation Research Part D Transport and Environment, vol. 154, p. 105265, Feb. 2026.

[11] G. De Nunzio, M. Laraki, and L. Thibault, "Road traffic Dynamic pollutant emissions estimation: From macroscopic road information to microscopic environmental impact," Atmosphere, vol. 12, no. 1, p. 53, Dec. 2020.

[12] E. Barmpounakis, M. Montesinos-Ferrer, E. J. Gonzales, and N. Geroliminis, "Empirical investigation of the emission-macroscopic fundamental diagram," Transportation Research Part D Transport and Environment, vol. 101, p. 103090, Oct. 2021.

[13] A. Gosztola, B. Verebélyi, and B. Horváth, "Traffic Modeling and Emission Calculation: Integration of the COPERT Method into the PTV-VISUM Software," Applied Sciences, vol. 16, no. 2, p. 567, Jan. 2026.

[14] S. K. Zegeye, B. De Schutter, J. Hellendoorn, E. A. Breunesse, and A. Hegyi, "Integrated macroscopic traffic flow, emission, and fuel consumption model for control purposes," Transportation Research Part C Emerging Technologies, vol. 31, pp. 158–171, Feb. 2013.

[15] Y.Q. Jiang, P.J. Ma, S.G. Zhou, "Macroscopic modeling approach to estimate traffic-related emissions in urban areas," Transportation Research Part D Transport and Environment, vol. 60, pp. 41-55, 2015.

[16] A. Csikós, T. Tettamanti, and I. Varga, "Macroscopic modeling and control of emission in urban road traffic networks," Transport, vol. 30, no. 2, pp. 152–161, May 2015.

[17] L. Heni, H. Haj-Salem, J.-P. Lebacque, and K. Slimi, "Integration and Comparative Analysis of COPERT and HBEFA Emission Models Coupled with the BIDIM-GSOM Traffic Model for Large-Scale Networks," Transportation Research Procedia, vol. 86, pp. 361–370, Jan. 2025.

[18] S. Kawsar, S. Biswas, M. Noor, and Md. S. Mamun, "Investigating the applicability of COPERT 5.5 emission software in Bangladesh and developing countrywide vehicular emission inventories," Environmental Science Atmospheres, vol. 4, no. 1, pp. 57–72, Nov. 2023.

[19] Ü. Ekici, S. Görk, and S. Moslem, "Predicting average annual daily traffic using machine learning: a comparative study for infrastructure investment," Transportation Research Interdisciplinary Perspectives, vol. 38, p. 102093, Jun. 2026.

[20] A. M. Ali, A. Al-Qerem, M. R. Hassan, A. Abu-Khadrah, and M. Jarrah, "Traffic monitoring and control system for smart city pollution regulation using IoT and correlated capsule networks," IEEE Access, vol. 13, pp. 46983–47003, Jan. 2025.

[21] S. Alamri and S. Khan, "Artificial intelligence-based modeling for predicting CO2 emission for climate change mitigation in Saudi Arabia," International Journal of Advanced Computer Science and Applications, vol. 14, no. 4, Jan. 2023.

[22] S. Dasgupta, D. Wheeler, M. Khaliquzzaman, and M. Huq, "Siting priorities for congestion-reducing projects in Dhaka: a spatiotemporal analysis of traffic congestion, travel times, air pollution, and exposure vulnerability," International Journal of Sustainable Transportation, vol. 16, no. 12, pp. 1078–1096, Aug. 2021.

[23] N. A. Policarpo, C. Silva, T. F. A. Lopes, F. S. A. Cavalcante, C. S. Pitombo, and R. S. Araújo, "Road vehicle emission inventory of a Brazilian metropolitan area and insights for other emerging economies," Transportation Research Part D Transport and Environment, vol. 58, pp. 172–185, 2017.

[24] A. K. Seyhan and N. Ertugay, "Traffic-related changes in air quality: An assessment based on PM10, NO2, SO2, and CO parameters," Gazi Üniversitesi Fen Bilimleri Dergisi Part C Tasarım Ve Teknoloji, vol. 13, no. 3, pp. 1100–1112, Aug. 2025.

[25] Z.-J. Ding, Z. Dai, X. Chen, and R. Jiang, "Simulating on-demand ride services in a Manhattan-like urban network considering traffic dynamics," Physica A Statistical Mechanics and Its Applications, vol. 545, p. 123621, Nov. 2019.

[26] A. M. S. Elarifi and W. M. S. Khalifa, "AI-Driven Predictive Analytics for Sustainable Aviation: Metaheuristic-Optimized XGBOOST for Carbon Emission Prediction," Sustainability, vol. 18, no. 5, p. 2246, Feb. 2026.

[27] L. Zhang, G. Lu, X. Yan, P. Xia, Z. Chen, and D. Wu, "A differential evolution optimized hybrid XGBoost for accurate carbon emission prediction," Environmental Modeling & Software, vol. 193, p. 106627, Jul. 2025.

[28] E. Aslan, "Comparison of predictions of CO2 emissions in vehicles using different artificial neural network models (Araçlarda CO2 emisyonlarının farklı yapay sinir ağı modelleri kullanılarak tahminlerinin karşılaştırılması)," Çukurova University Journal of the Faculty of Engineering, vol. 39, no. 2, pp. 309–324, Jul. 2024.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Multi-Source Behavior-Based Anomaly Detection for Smart Water Management Networks Using SNMP and Syslog Telemetry

Dilara ÇETİN
Dept. of Computer Engineering
Sakarya University
Sakarya, Türkiye
dilaracetin13@gmail.com

Ayşegül KARA
Dept. of Computer Engineering
Sakarya University
Sakarya, Türkiye
aysegulkaraa3@gmail.com

Deniz BALTA
Dept. of Software Engineering
Sakarya University
Sakarya, Türkiye
ddural@sakarya.edu.tr

Musa BALTA
Dept. of Cyber Security Engineering
Sakarya University
Sakarya, Türkiye
mbalta@sakarya.edu.tr

Abstract

Water management systems are critical infrastructure that are increasingly exposed to cyberattacks and failure risks due to their growing connectivity to corporate networks and the Internet. In this study, an AI-based real-time anomaly detection system was developed to monitor the SCADA/OT network of a water management facility. The study was conducted at the National Critical Infrastructure Testbed Center at Sakarya University. The system collects SNMP, ping, and syslog data from 28 devices, transforms them into a 224-channel multivariate time series, and compares GDN, TranAD, LSTM-AE, and Anomaly Transformer models under the same experimental conditions. Due to the limited availability of real anomaly data, a hybrid synthetic data generation method that preserves the statistical structure of normal traffic was used, and the generated data were validated using the Kolmogorov-Smirnov test, PCA, and the TSTR (Train-on-Synthetic, Test-on-Real) method. Experimental results show that GDN achieved the best performance on hard-level anomalies, with ROC-AUC of 0.934, F1 of 0.979, and PR-AUC of 0.802. The results are presented through a web-based monitoring dashboard that generates explainable alarms enriched with MITRE ATT&CK for ICS mappings.

Keywords: Smart water management, cyber security, anomaly detection, deep learning

I. INTRODUCTION

Water management facilities are critical infrastructures that directly affect public life through processes such as the treatment, pumping, and storage of drinking water [1]. Failures or cyberattacks in these systems may lead to physical consequences, including water outages, service disruptions, and adverse effects on public health. Therefore, the security of SCADA/OT systems should be addressed through a holistic approach that covers not only information security but also the continuity of physical processes [1], [2]. Although SCADA/OT systems were traditionally operated largely in isolation from external networks, they are increasingly connected to corporate networks and the Internet due to remote access, centralized monitoring, and operational efficiency requirements. This expansion increases the attack surface and makes critical infrastructures more vulnerable to cyber threats. Incidents such as Stuxnet and the Maroochy Shire breach have demonstrated that attacks targeting industrial control systems can cause physical and environmental consequences [3], [4]. Therefore, continuous monitoring and early anomaly detection are required in addition to access control and network security measures. Traditional monitoring tools mostly focus on the availability of OT assets, basic counter values, and system logs. However, attacks in OT networks often emerge as low-magnitude, time-distributed behavioral changes that leave

simultaneous traces across multiple devices. Therefore, rather than only monitoring device status, methods that can learn the network's normal behavior and detect deviations early are needed.

One of the main challenges in this field is the limited availability of real anomaly data. Collecting attack or failure data in critical infrastructure environments is both difficult and risky. Moreover, such events are rare, which prevents the collection of sufficient samples for training deep learning models. This limitation necessitates the generation of controllable, labeled synthetic data that preserves the structure of real normal traffic.

In this study, an AI-based real-time anomaly detection system was developed for SCADA/OT networks in water management systems. SNMP, ping, and syslog data were collected from 28 devices and transformed into a 224-channel multivariate time-series representation. GDN, TranAD, LSTM-AE, and Anomaly Transformer models were compared using the same data representation and evaluation approach. Due to the limited availability of real anomaly samples, hybrid synthetic data generation was applied, and the models were evaluated under easy, medium, and hard difficulty levels. The developed system continuously scores the collected data and notifies the operator of potential anomalies through a web-based monitoring dashboard.

The main contributions of this study can be summarized as follows:

- A labeled and reproducible hybrid synthetic data generation approach was developed to preserve the statistical structure and fundamental behavioral relationships of real normal traffic. This provides a controlled data generation process for model training and evaluation in OT/ICS environments where real anomaly data are limited.
- Data obtained from SNMP, ping, and syslog sources were transformed into device-based features, resulting in a 224-channel multivariate time-series representation over 28 devices. In this way, the study aims to model system-wide relational behavioral patterns rather than isolated device-level signals.
- GDN, TranAD, LSTM-AE, and Anomaly Transformer models were compared under the same data representation, preprocessing pipeline, and evaluation metrics. This comparison provides a fair experimental framework for analyzing the performance of different model families in SCADA/OT anomaly detection.
- Anomalies were generated at three different difficulty levels: easy, medium, and hard. This enabled evaluation of model behavior not only against obvious anomalies but also against low-magnitude, difficult-to-detect events.
- The outputs of the developed anomaly detection system were integrated into a real-time web-based monitoring dashboard. Alarm records were enriched with risk levels, explanatory fields, and MITRE ATT&CK mappings for ICS, making them more interpretable for operators.

II. LITERATURE REVIEW AND TECHNICAL BACKGROUND

Anomaly detection methods are generally classified into statistical approaches, classical machine learning methods, and deep learning-based models. While statistical and classical methods rely on specific assumptions and predefined features, the multidimensional and temporal structure of SCADA/OT data highlights the need for deep learning models capable of learning complex patterns [5]. An important group of deep learning-based approaches consists of reconstruction-based methods. These models learn to reconstruct normal data and use the reconstruction error as an anomaly score. LSTM-Autoencoder is a common example of this approach, as it can model long-term temporal dependencies [6]. In prediction-based methods, future values are estimated from past observations, and the prediction error is evaluated as an anomaly indicator. GDN provides an advantage in detecting correlated anomalies by learning the relationships between devices through a graph structure [7]. Transformer-based models can capture long-term dependencies and contextual inconsistencies through the attention mechanism. While TranAD aims to detect small deviations using a two-stage reconstruction structure

[8], Anomaly Transformer identifies contextual anomalies through an association discrepancy measure [9]. In this study, these four models were compared using the same data representation and evaluation approach.

Although GNN- and Transformer-based methods have achieved strong results in OT/ICS anomaly detection, most existing studies focus primarily on model performance on offline datasets. Real-time monitoring, explainable alarm generation, and operator support remain limited. This study contributes to the literature by integrating the anomaly detection model into an end-to-end monitoring architecture and presenting model outputs together with risk level, affected protocol, and MITRE ATT&CK for ICS mapping. Open datasets such as UNSW-NB15 [10], CIC-IDS2017 [11], CIC-DDoS2019 [12], InSDN [13], and SWaT [14], [15] are widely used in anomaly detection studies. However, UNSW-NB15, CIC-IDS2017, and CIC-DDoS2019 mainly focus on general IT network traffic, whereas InSDN focuses on SDN environments. Although SWaT is more closely related to this study as an ICS dataset based on a water treatment process, it does not directly represent the multi-source OT monitoring structure used in this work, which consists of SNMP counters, syslog records, and ping measurements. Therefore, labeled and controllable synthetic data generation was preferred to fit the targeted architecture.

III. PROPOSED MULTI-SOURCE BEHAVIORAL ANOMALY DETECTION FRAMEWORK

This section describes the data collection process and the synthetic data generation approach that form the basis of the anomaly detection system developed in this study. Within the scope of the study, SNMP, syslog, and ping data were collected from SCADA/OT components in the laboratory environment. Since real anomaly samples are limited, a hybrid synthetic data generation method was applied to preserve the statistical structure of these data. In this way, the normal traffic behavior that reflects the real environment was preserved, while different attack and failure scenarios were presented to the model in a controlled manner.

1. Testbed Environment and Monitoring Assets

The data used in this study were collected in a laboratory environment that represents a water management system at the National Critical Infrastructure Testbed Center of Sakarya University. The laboratory network consists of pumping, treatment, and storage stations, and the devices are separated into IP-based subnets in accordance with the zone-and-conduit approach of the IEC 62443 standard [2]. Thus, devices with similar security requirements are grouped within the same zone, while communication between zones is provided through controlled conduits. A total of 28 devices, including PLCs, switches, HMIs, and firewalls, are monitored in the environment. PLCs control field equipment, switches forward network traffic, HMIs provide the operator interface, and firewalls regulate inter-station and external traffic. The collector component, running in the background, queries the devices every 15 seconds and records the data in a central SQLite database. SNMP, syslog, and ping sources are used during data collection. Through SNMP, counter values for packets, errors, responses, and management requests are obtained; cumulative counters are computed by taking the difference between consecutive measurements [16]. Delta values of ten different SNMP counters are collected from each device. Syslog provides event, status, and error messages with timestamp, source, facility, and severity fields [17]. Ping generates information on device availability, latency, and packet loss [18]. These three sources are used together to monitor device status and network behavior holistically.

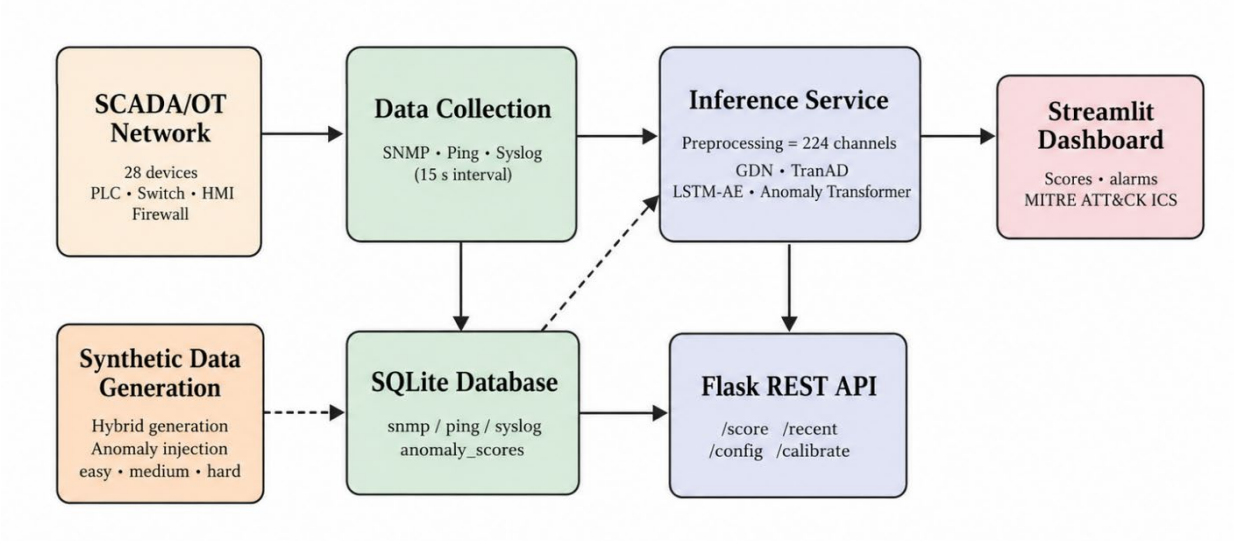


Figure 1. Real-Time SCADA/OT Anomaly Monitoring Pipeline

2. Extraction of Structural Invariants

To generate realistic synthetic data, the structural invariants of normal traffic were first identified. For this purpose, SNMP counter distributions, ratios between counters, and temporal behaviors in the real data were analyzed. As a result of this analysis, four main relationships were observed: a packet balance exists among InPkts, OutPkts, and OutGetResponses; GetNexts requests constitute approximately 92% of total packets; BadCommunityNames, SetRequests, and ASN error counters remain zero under normal conditions; and counters are updated at regular intervals of approximately 74 seconds. Preserving these invariants is important for ensuring that the synthetic data remains consistent with the real environment. Otherwise, the model may learn inconsistent patterns that would not occur in the real system. Therefore, packet balance, counter ratios, and normal error behavior were preserved during the generation process.

3. Hybrid Synthetic Data Generation

To address the class imbalance problem, this study employs a transparent and reproducible hybrid approach to synthetic data generation that preserves the statistical structure of real data. Although learning-based methods such as GANs can generate realistic data, their training processes are unstable, and their explainability is limited [19]. Therefore, this study combines baseline generation calibrated with measured statistics of normal traffic and anomaly injection based on documented attack rules. The generation process consists of two stages. In the first stage, an anomaly-free normal baseline is generated for each device using the average packet rate measured from real data. To reflect the variability of real network traffic, a time-of-day load pattern and device-specific log-normal noise are added to this value. Structural invariants are preserved in each measurement; outgoing packets are balanced with incoming packets, the GetNexts ratio is adjusted to remain approximately 92%, and error counters are kept at zero under normal conditions. In the second stage, anomalies are injected into this baseline in a controlled manner.

4. Definition and Generation of Anomaly Types

Anomalies were added to the generated normal baseline according to a threat taxonomy suitable for critical infrastructure environments. The anomalies were not generated randomly; instead, they were based on the traces

left by real attack and failure scenarios in SNMP counters. In this context, seven different anomaly types were defined. For example, network reconnaissance increases GetNexts values, represented as walk_spike; unauthorized configuration attempts increase the SetRequests counter, represented as set_attempt; and DoS attacks increase InPkts values while disrupting the response/request balance, represented as flood. The anomaly types are summarized in Table 1.

Table 1. Injected Anomaly Types And Scenarios

Anomaly Type	**Counter Trace (Signature)**	**Corresponding Scenario**
walk_spike	GetNexts suddenly increases	Network reconnaissance
set_attempt	SetRequests rises from zero	Unauthorized configuration write
community_brute	BadCommunityNames increases	Community name brute-force attempts
flood	InPkts surges; responses < requests	Denial of Service (DoS)
silent_device	All counters are reset to zero	Device shutdown/connectivity loss
asn_errors	ASN parsing errors increase	Malformed/malicious packet
response_drop	Response count falls below requests	Overloaded/unresponsive device

Each anomaly is defined by severity and duration parameters. In this way, different difficulty levels can be generated for the same anomaly type. The injected anomalies are added without overlap and labeled for use during the evaluation stage. In addition, SNMP, ping, and syslog data are generated in a correlated manner rather than independently, so that they represent different reflections of the same event. For example, a device failure is modeled simultaneously through the reset of SNMP counters, ping unavailability, and a syslog error record. To evaluate model performance more realistically, anomalies are generated at three difficulty levels: easy, medium, and hard. At the easy level, large and long-lasting deviations are generated. At the medium level, lower-magnitude anomalies are produced. At the hard level, small, gradual deviations from normal behavior are introduced. This difficulty gradient makes performance differences between models more apparent.

5. Validation of Synthetic Data

The representativeness of the synthetic data with respect to real traffic was validated statistically and functionally. In the statistical validation, structural ratios were first checked; the outgoing/incoming packet ratio was verified to be 1.0, and the SNMP walk ratio was confirmed to remain around 92%. Then, the compatibility of counter distributions with real data was evaluated using the Kolmogorov-Smirnov (KS) test [20]. In the tests conducted for each SNMP counter, the synthetic distributions did not differ significantly from the real distributions. In addition, principal component analysis (PCA) showed that synthetic normal data were concentrated in the same region as real normal data [21]. For functional validation, the TSTR (Train-on-Synthetic, Test-on-Real) method was used. The model was trained only on synthetic data and tested on real laboratory data unseen during training. The ability of the model trained on synthetic data to distinguish limited anomalies in real data from normal traffic indicates that the generated data are also suitable for learning in real environments.

6. System-Wide Preprocessing and Feature Extraction

Since deep learning models require numerical inputs sampled at equal time intervals, SNMP, ping, and syslog data generated at different periods were resampled onto a common one-minute time grid. Text-based syslog records were not used directly; instead, they were converted into numerical indicators, such as the number of messages by severity level within the corresponding time window. In this study, devices are evaluated not individually but through the simultaneous behavior of the entire network. Eight features are extracted from each device, resulting in a $28 \times 8 = 224$-dimensional feature vector at each time step for 28 devices. This approach was preferred because OT/ICS attacks may leave low-magnitude and simultaneous traces across multiple devices. All features were scaled to the 0–1 range. Normalization parameters were computed only from the training set and then applied to the test set using the same values, thereby preventing information leakage.

7. Models Used

In this study, LSTM-AE, TranAD, Anomaly Transformer, and GDN models were compared. All models were trained only on normal data using a semi-supervised approach, and the high reconstruction or prediction error produced for inputs deviating from normal behavior was used as the anomaly score. LSTM-AE was selected as the baseline model, while TranAD was evaluated for its two-stage Transformer architecture, designed to capture small deviations. Anomaly Transformer analyzes contextual inconsistencies using the association discrepancy measure. In contrast, GDN learns relationships between features through a graph structure, making it a strong candidate for multi-device OT/ICS environments.

8. Training, Testing, and Real-Time Threshold Calibration

The models were trained only on normal data, using a 48-hour training set that contained no anomalies. During the test phase, a fully independent dataset containing labeled anomalies was used. To ensure a methodologically fair comparison, all four models were evaluated using the same training set, test sets, and evaluation metrics. Thus, the observed performance differences can be attributed to the models' capabilities rather than to differences in experimental conditions. The trained models were integrated into a real-time monitoring pipeline. At the end of each data collection cycle, a scoring process is triggered; the most recent data are retrieved, processed through the same preprocessing steps, scored for the latest window, and recorded in the database. The system initially operates in calibration mode. In this mode, scores are calculated and stored, but no alarm is generated. This is because generating alarms before determining a meaningful threshold may lead to a high false-positive rate. The threshold value is learned from the system's real normal traffic. During a defined observation period, the distribution of normal scores is analyzed, and the threshold is set at an upper percentile of this distribution, such as 99.5%. The calibration process automatically removes unreliable warm-up scores at the beginning; otherwise, the threshold may be set higher than necessary, causing real anomalies to be missed. Since the threshold is stored in the database rather than hard-coded and can be updated through the interface, the system can be recalibrated without code changes when it is transferred to a different environment. This design choice is also consistent with the ISO/IEC 27001 principles of continuous monitoring and improvement [22].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

1. Evaluation Metrics

ROC-AUC, PR-AUC, and point-adjusted F1 metrics were used to evaluate anomaly detection performance. ROC-AUC measures a model's ability to distinguish between normal and anomalous samples independently of the decision threshold [23]. PR-AUC is particularly important for imbalanced datasets in which anomalies are rare, as it also reflects the tendency of false positives [24], [25]. Point-adjusted F1 is a metric commonly used in

time-series anomaly detection, where an anomaly event is considered detected if any point within the event is correctly identified [26]. To improve the reliability of the results, each experiment was repeated with three different random seeds, and the average performance values were reported.

2. Results by Difficulty Level

The ROC-AUC values for the four models across three difficulty levels, along with their F1 performance at the hard level, are summarized in Table 2. At the easy and medium levels, all models produced high and similar ROC-AUC values; therefore, the main performance differences became evident at the hard level. At the hard level, GDN achieved the highest ROC-AUC value of 0.934, outperforming TranAD (0.833) and Anomaly Transformer (0.835), while LSTM-AE showed the lowest performance with 0.773 (Figure 2). The F1 results align with this trend, indicating that GDN can detect hard anomalies with high precision and recall, achieving an F1 score of 0.979.

Table 2. Performance Metrics By Difficulty Level

Model	**Easy ROC**	**Medium ROC**	**Hard ROC**	**Hard F1**
LSTM-AE	0.972	0.955	0.773	0.697
TranAD	0.998	0.990	0.833	0.955
Anomaly Transformer	0.999	0.991	0.835	0.968
GDN	0.995	0.989	0.934	0.979

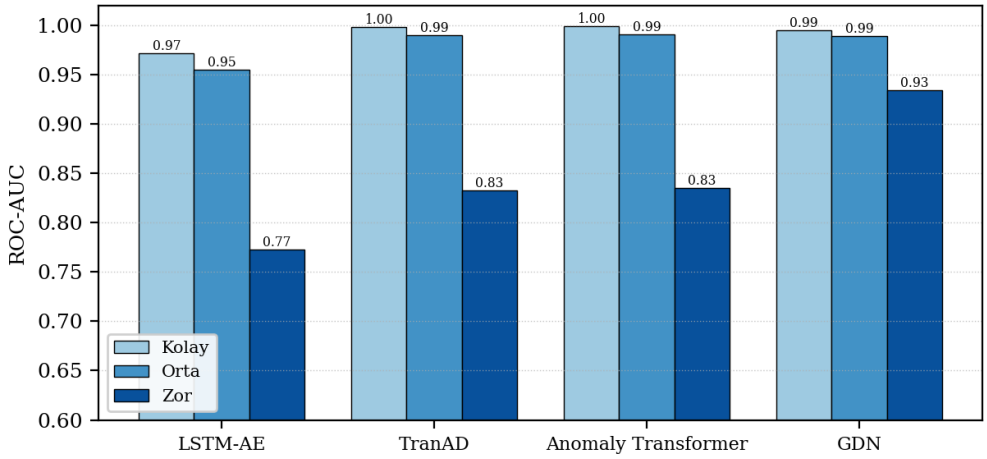


Figure 2. Performance metrics of the models by anomaly difficulty level.

3. Precision, Recall, and Stability Analysis

The practical value of an anomaly detection system depends on its ability to limit false positives without missing real anomalies. In the hard-level PR-AUC results, GDN achieved the most balanced performance with a score of 0.802, while LSTM-AE showed the lowest performance with 0.411. TranAD and Anomaly Transformer

achieved comparable results, with PR-AUC values of 0.697 and 0.707, respectively. In the multi-seed experiments, GDN also maintained high average performance with low standard deviation. These results indicate that GDN is a more stable candidate for deployment across different data periods or different OT/ICS environments.

V. CONCLUSION

This paper presented a behavior-based anomaly detection framework for smart water management SCADA/OT networks using multi-source monitoring data collected from SNMP, syslog, and ping telemetry. To overcome the limited availability of real anomaly data in critical infrastructure environments, a hybrid synthetic data generation approach was developed to preserve the statistical characteristics and structural invariants of normal network traffic while enabling the controlled injection of representative attack and failure scenarios. Four state-of-the-art deep learning models, namely LSTM-AE, TranAD, Anomaly Transformer, and GDN, were evaluated under identical preprocessing and experimental conditions. The experimental results demonstrated that the graph-based GDN model consistently outperformed the other approaches, particularly for hard-to-detect anomalies, achieving an ROC-AUC of **0.934**, an F1-score of **0.979**, and a PR-AUC of **0.802**, indicating its superior capability to capture correlated behavioral deviations across multiple OT devices. Beyond offline model evaluation, the proposed framework integrates multi-source data acquisition, feature extraction, real-time anomaly scoring, adaptive threshold calibration, and explainable alarm visualization into a unified monitoring architecture. These findings demonstrate that combining behavioral analysis with multi-source network telemetry provides an effective and practical solution for improving cyber resilience and operational awareness in smart water management infrastructures. Future work will focus on validating the framework using real-world cyberattack scenarios, incorporating additional OT communication protocols and telemetry sources, and investigating online learning techniques to enable adaptive anomaly detection in continuously evolving operational environments.

REFERENCES

[1] K. Stouffer et al., Guide to Operational Technology (OT) Security, NIST Special Publication 800-82 Revision 3. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2023.

[2] International Electrotechnical Commission, Security for Industrial Automation and Control Systems—Part 3-2: Security Risk Assessment for System Design, IEC 62443-3-2:2020, 2020.

[3] R. Langner, “Stuxnet: dissecting a cyberwarfare weapon,” IEEE Security & Privacy, vol. 9, no. 3, pp. 49–51, 2011.

[4] J. Slay and M. Miller, “Lessons learned from the Maroochy water breach,” in Critical Infrastructure Protection, IFIP Advances in Information and Communication Technology, vol. 253, E. Goetz and S. Shenoi, Eds. Boston, MA, USA: Springer, 2008, pp. 73–82.

[5] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: a survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[6] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, “LSTM-based encoder-decoder for multi-sensor anomaly detection,” arXiv:1607.00148, 2016.

[7] A. Deng and B. Hooi, “Graph neural network-based anomaly detection in multivariate time series,” in Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 5, pp. 4027–4035, 2021.

[8] S. Tuli, G. Casale, and N. R. Jennings, “TranAD: deep Transformer networks for anomaly detection in multivariate time series data,” Proc. VLDB Endowment, vol. 15, no. 6, pp. 1201–1214, 2022.

[9] J. Xu, H. Wu, J. Wang, and M. Long, “Anomaly Transformer: time series anomaly detection with association discrepancy,” in Proc. Int. Conf. Learning Representations, 2022.

[10] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems," in Proc. Military Communications and Information Systems Conf., pp. 1–6, 2015.
[11] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. Int. Conf. Information Systems Security and Privacy, pp. 108–116, 2018.
[12] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in Proc. Int. Carnahan Conf. Security Technology, pp. 1–8, 2019.
[13] M. S. Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, "InSDN: a novel SDN intrusion dataset," IEEE Access, vol. 8, pp. 165263–165284, 2020.
[14] A. P. Mathur and N. O. Tippenhauer, "SWaT: a water treatment testbed for research and training on ICS security," in Proc. Int. Workshop Cyber-Physical Systems for Smart Water Networks, pp. 31–36, 2016.
[15] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in Critical Information Infrastructures Security, LNCS, vol. 10242. Cham, Switzerland: Springer, 2017, pp. 88–99.
[16] D. Harrington, R. Presuhn, and B. Wijnen, An Architecture for Describing SNMP Management Frameworks, RFC 3411, IETF, 2002.
[17] R. Gerhards, The Syslog Protocol, RFC 5424, IETF, 2009.
[18] J. Postel, Internet Control Message Protocol, RFC 792, IETF, 1981.
[19] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued medical time series generation with recurrent conditional GANs," arXiv:1706.02633, 2017.
[20] F. J. Massey Jr., "The Kolmogorov-Smirnov test for goodness of fit," Journal of the American Statistical Association, vol. 46, no. 253, pp. 68–78, 1951.
[21] I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002.
[22] ISO/IEC 27001:2022, Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements, ISO/IEC, 2022.
[23] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[24] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015.
[25] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. Int. Conf. Machine Learning, pp. 233–240, 2006.
[26] H. Xu et al., "Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications," in Proc—the Web Conf., pp. 187–196, 2018.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Novel PL/SQL-Based Data Transformation and Optimization Framework for Enterprise Data Warehouses

Tirumalavenkata Naga Lakshmi Jammula
Independent Researcher
Charlotte, NC, United States of America
lak.jammula@gmail.com
ORCID: 0009-0004-7582-2524

Abstract
With the rapid growth of enterprise data, significant challenges have emerged in managing, transforming, and optimizing the datasets collected in an Enterprise Data Warehouse (EDW). Traditional Extract, Transform, and Load (ETL) methods rely extensively on external tools, resulting in operational inefficiencies such as increased data transfer, higher latency, increased operational costs, and underutilization of available resources. This paper proposes a novel PL/SQL-based data transformation and optimization framework that leverages Oracle Database's in-database processing capabilities. The framework integrates advanced PL/SQL operations, such as bulk collection and parallelism, and utilizes various database-level optimization techniques, including index creation and partitioning. The experimental results demonstrate improved efficiency in terms of execution time and fault tolerance.

Keywords: PL/SQL, Data Warehousing, ETL Optimization, Oracle Database, Data Transformation, Performance Tuning, Metadata-Driven Framework

I. INTRODUCTION

In the contemporary IT environment, companies generate vast volumes of data from various sources, including transaction systems, websites, mobile platforms, and external data sources. Effective management and transformation of large volumes of data are essential to ensure successful business intelligence activities. The Enterprise Data Warehouse (EDW) serves as the central repository for data from different sources and as the basis for reporting. As data volumes grow and become increasingly complex, data transformation processes are becoming more difficult. The extraction, transformation, and loading (ETL) process is a major aspect of the data warehousing solution. It involves obtaining unstructured data from various sources, applying transformation rules, and loading the transformed data into the warehouse tables. Contemporary ETL approaches commonly rely on third-party software solutions such as Informatica PowerCenter and Microsoft SQL Server Integration Services. These solutions enable effective integration and data management; however, they are accompanied by several limitations because data must be moved across different platforms. Data migration requires additional resources, including bandwidth and computing power, which reduces efficiency. To resolve this problem, in-database processing has proven highly effective in improving ETL process efficiency. In-database processing involves performing transformations within the database server rather than moving data to another location to apply transformation logic. Oracle DBMS offers one of the most popular procedural languages for SQL processing: PL/SQL. Thus, it becomes possible to apply business rules, perform transformations, and perform batch processing within the database. Several advantages are associated with using PL/SQL code, including opportunities for large-scale data manipulation. For example, it is possible to use BULK COLLECT and FORALL operations to perform transformations faster, since there is no context switching between the SQL and procedural execution engines. There are numerous additional features in PL/SQL, including support for procedures and packages, dynamic SQL, exception handling, and more. However, limited enterprise adoption of PL/SQL is systematic. This paper proposes a novel ETL approach that uses PL/SQL data transformations and optimizations. The system architecture includes components such as a data ingest layer, a data transform layer,

a data optimization layer, and a data storage layer. Advanced PL/SQL techniques and database optimizations are used. Additionally, the system includes components such as error handling, logging, and monitoring. The Novelty of this paper lies in the creation of a comprehensive, reusable, and metadata-driven transformation framework for enterprise data warehouse workloads using PL/SQL bulk processing, set-based transformation, database optimization, logging, and a restartable, fault-tolerant architecture.

In conclusion, this framework contributes the following:

- It presents an ETL/ELT framework driven by metadata based on PL/SQL for enterprise data warehouses.
- It provides integration of staging, validation, transformation, optimization, auditing, logging, and error recovery within Oracle Database.
- It tests the ETL/ELT framework by conducting performance evaluation with EDW benchmarking workload of a banking nature with high transaction rates.
- It compares the proposed ETL/ELT framework with the traditional ETL frameworks in terms of performance measures.

II.LITERATURE REVIEW

The Extract, Transform, Load (ETL) process is an essential part of data warehouse implementation in enterprises, providing mechanisms to integrate data from various sources into a single repository for analysis. As data volumes grow, the importance of effective ETL processes has increased. Data extraction, transformation into more meaningful forms, and loading of transformed data into warehouses are the main functions of the ETL pipeline [1], [2]. Significant changes have occurred in the ETL domain since the initial development of such systems. Earlier versions operated in batch mode and needed considerable human input to function effectively. Contemporary ETL systems offer automated operations, distributed computing, and cloud-based infrastructure. The ELT (Extract, Load, Transform) concept moved the transformation phase into a database environment [3]. Kimball and Caserta stressed the resource-intensive nature of the ETL phase in data warehousing and the need for proper optimization [4]. The same was argued by Vassiliadis et al., who noted the complexity of the process and the necessity of formalized modeling techniques [5]. The topic of ETL optimization has attracted considerable attention, and many methods have been suggested to optimize ETL process performance. For instance, Liu developed techniques such as shared coaching and parallelization, which can greatly reduce execution time and resource utilization for ETL processes [6]. These techniques help avoid unnecessary computations and effectively utilize system resources. Simitsis et al. indicated that techniques such as task scheduling optimization, data partitioning, and pipelining could significantly improve ETL performance in large organizations [7]. Such optimizations play a critical role in handling big data processes in corporate organizations. Push-down optimization can be considered a database-oriented approach in ETL architecture, as ETL logic is executed on the database side rather than by other processing systems. Kodali et al. reported that push-down optimizations significantly reduce data movement and speed up operations in a multi-cloud architecture [8]. Yu et al. demonstrated that moving computation closer to storage systems greatly improves query performance and cost efficiency by a factor of 6.7 [9]. Zhang et al. introduced automated predicate pushdowns, which significantly optimize the query execution process [10]. These findings confirm the efficiency of the database-oriented methodology and are consistent with PL/SQL transformation models. Automation is now an essential aspect of any current ETL system. Automation helps eliminate manual processes and improve performance scalability. Ebadifard et al. [11] highlighted the significance of automation in ETL pipelines for machine learning. Modern ETL solutions feature metadata management and intelligent scheduling, enabling performance optimization and real-time operations. However, existing ETL systems are subject to drawbacks such as long latency caused by data transport, Inability to scale under high load, costly operations and

maintenance, and Inflexibility for advanced transformations. Such problems are associated with the use of external engines rather than native database capabilities [12]. PL/SQL provides a robust mechanism for executing transformations within Oracle databases. As noted by Feuerstein and Pribyl, PL/SQL improves efficiency by reducing context switches between the application and database layers [13]. Harrison underlines the importance of improving performance at the database level in enterprise applications [14]. Recent literature on ETL methodologies, tools, processes, and evaluations further highlights the importance of effective transformations and sound architectural decisions for handling large volumes of data in a data warehouse [15]. At the same time, comprehensive ETL assessment emphasizes scalability and tool choice [16]. Pipeline quality analysis stresses validation and reliability issues [17]. Database benchmarking provides the basis for system performance evaluation [18], and Oracle SQL error management ensures fault tolerance [19]. The database-centered ETL approach using PL/SQL reduces transfer time, speeds up the process, and ensures better data consistency. Despite numerous studies on ETL optimization and automation, a coherent framework that incorporates ETL optimization techniques in PL/SQL environments remains lacking. Current approaches rely on separate ETL tools. This research aims to fill the gap mentioned above by developing a single PL/SQL framework that integrates optimization techniques, such as parallel processing and pushdown execution, into transformation algorithms.

III. FRAMEWORK ARCHITECTURE

The architecture is designed as a multi-layered system that encompasses processes such as data intake, transformation, optimization, and storage. This framework leverages Oracle Database's database functionality to optimize the process while minimizing data transport. Each layer is associated with a specific ETL process.

1. Data Ingestion Layer

The Data Ingestion Layer involves extracting data from sources such as transactional databases, flat files, and other external data stores. It uses software such as SQL*Loader and external tables to rapidly load data into staging areas. The external tables feature enables users to query data files directly without importing them into the database. At this level, preliminary checks are carried out to ensure data integrity. Some checks include null values, data type conflicts, and data format issues. The cleaned data is subsequently loaded into staging tables.

2. Transformation Engine (PL/SQL Core)

The Transformation Engine is the central component of the proposed framework and is implemented in PL/SQL in Oracle Database. The Transformation Layer transforms raw data and applies business logic to produce meaningful, analyzable data. The use of stored procedures and packaged code improves maintainability and promotes code reusability. Techniques such as bulk collect and for all are used to handle high volumes of data by minimizing context switching between the SQL and PL/SQL engines. Also, this layer cleanses, enriches, aggregates, and normalizes data according to the business rules. Complex transformations are performed in the database, including multi-table joins and derived attributes.

3. Optimization Layer

The Optimization Layer is a critical component for enhancing the performance of the transformation process. This layer employs multiple database-level optimization techniques to ensure efficient processing of large-scale datasets. Indexing techniques, including B-tree indexing for high-cardinality columns and Bitmap indexing for low-cardinality columns, are used to improve query performance. Table partitioning is another optimization

technique that divides large tables into smaller, more manageable segments, enabling partition pruning. Parallel execution distributes workloads across multiple CPUs, enabling efficient execution of computationally intensive tasks. Query tuning helps identify and eliminate inefficiencies, including unnecessary full-table scans.

4. Data Storage Layer

The Data Storage Layer stores processed data in the company's enterprise data warehouse. The organization of the data in this layer follows dimensional models such as the star or snowflake schema, which are common approaches for performing analytical processing. Specifically, the layer uses fact tables for quantitative data, such as transactions or measurements, whereas dimension tables hold descriptive information about customers, products, or time periods. The well-structured nature of this layer enables efficient queries and reports that deliver business intelligence. Moreover, performance can be further improved through indexing and aggregation techniques.

In summary, the described approach incorporates all four layers, creating an effective solution that enhances ETL efficiency, minimizes data processing time, and improves scalability through in-database processing.

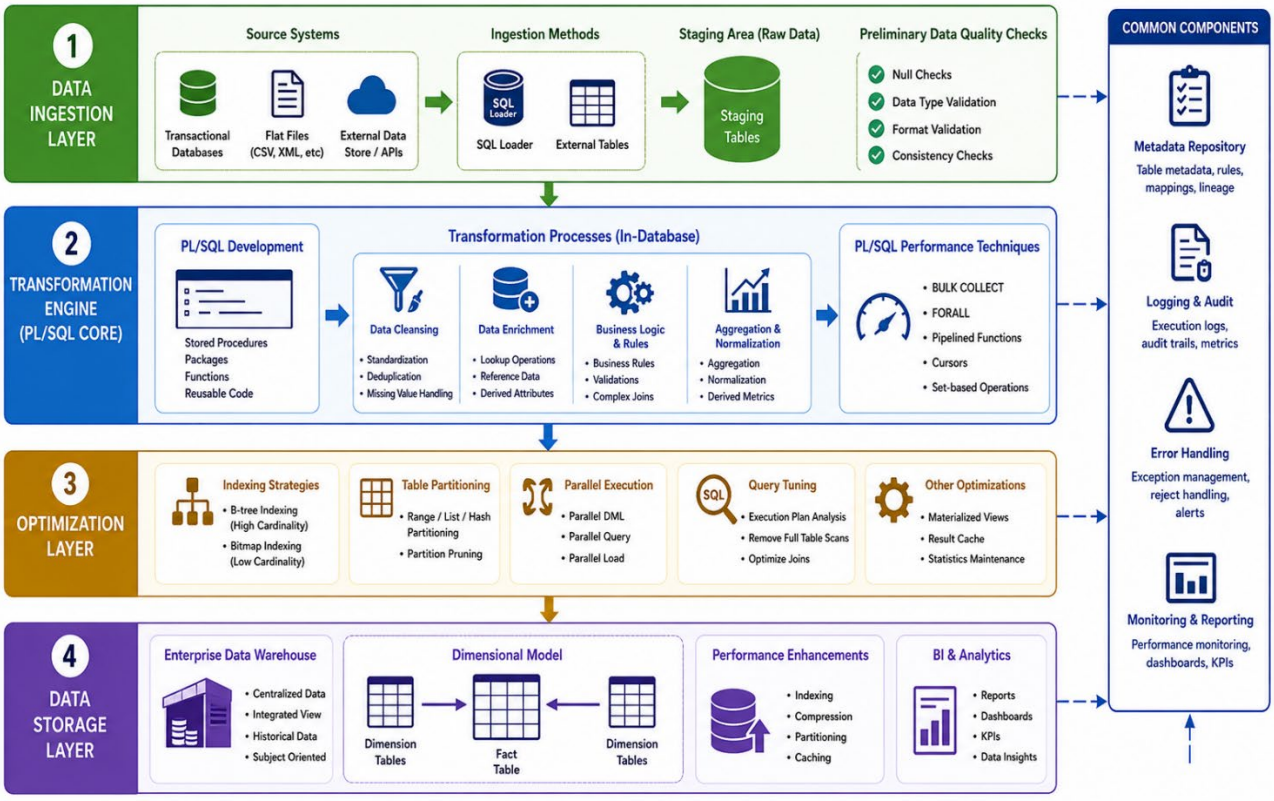


Figure 1. Architecture of a PL/SQL-Based Data Transformation and Optimization Framework for Enterprise Data Warehouses

IV.METHODOLOGY



This study proposes applying an ETL solution based on an in-database technique that leverages PL/SQL capabilities to improve data transformation performance. Data acquisition is the initial stage that includes loading data from diverse sources, such as transactional systems, flat files, and external data storage systems. The SQL*Loader tool and external tables are used to load data into the staging area for preliminary validation, including checking for nulls, data types, and formats.

Subsequently, PL/SQL transformation engines transform the data uploaded to the staging area using stored procedures, functions, and packages that provide data cleaning and validation, data enrichment, and business rule implementation. Techniques such as BULK COLLECT and FORALL are used to process large volumes of data and minimize context switches. Optimization of data transformation processes includes implementing indexing, table partitioning, parallel processing, and execution-plan-based query tuning. The last stage involves loading data into the enterprise data warehouse and dimensional modeling of the data structure (star or snowflake schema). In addition, metadata management systems, logging, and monitoring processes guarantee efficient data quality management. Unlike traditional ETL tools that operate externally, the proposed framework minimizes unnecessary movement of large datasets from the database to external transformation engines. Unlike traditional Oracle ELT scripts, the presented framework is not limited to isolated SQL procedures. Instead, it offers a structured design with metadata-driven rule execution, reusable PL/SQL packages, audit trails, record-rejection capabilities, checkpointing for restart, and database-level optimization.

1. BULK COLLECT with Error Isolation

A critical design consideration for the BPE is the use of the SAVE EXCEPTIONS clause with FORALL DML statements, accompanied by an effective exception handler that can track each row-level error without halting the entire process. In this manner, the throughput is optimized for valid records while maintaining comprehensive traceability for invalid records.

```
PROCEDURE load_staging_bulk(  p_pipeline_id IN NUMBER,  p_batch_size  IN PLS_INTEGER
DEFAULT NULL) IS
  TYPE t_source_tab IS TABLE OF stg_raw_data%ROWTYPE;
  v_source_data  t_source_tab;
  v_batch_size   PLS_INTEGER;
  v_errors       PLS_INTEGER := 0;
  ex_dml_errors  EXCEPTION;
  PRAGMA         EXCEPTION_INIT(ex_dml_errors, -24381);
  CURSOR c_source IS
    SELECT * FROM stg_raw_data
    WHERE  pipeline_id = p_pipeline_id
    AND    status = 'PENDING';
BEGIN
  v_batch_size := NVL(p_batch_size,
    compute_optimal_batch_size('STG_RAW_DATA'));
  OPEN c_source;
  LOOP
    FETCH c_source BULK COLLECT INTO v_source_data
    LIMIT v_batch_size;
    EXIT WHEN v_source_data.COUNT = 0;
    BEGIN
      FORALL i IN 1..v_source_data.COUNT SAVE EXCEPTIONS
        INSERT INTO stg_cleansed_data VALUES v_source_data(i);
    EXCEPTION
```

```
    WHEN ex_dml_errors THEN
     FOR j IN 1..SQL%BULK_EXCEPTIONS.COUNT LOOP
      log_row_error(
       p_pipeline_id, v_source_data(SQL%BULK_EXCEPTIONS(j).ERROR_INDEX),
       SQL%BULK_EXCEPTIONS(j).ERROR_CODE
      );
      v_errors := v_errors + 1;
     END LOOP;
    END;
  END LOOP;
  CLOSE c_source;
END load_staging_bulk;
```

V.IMPLEMENTATION

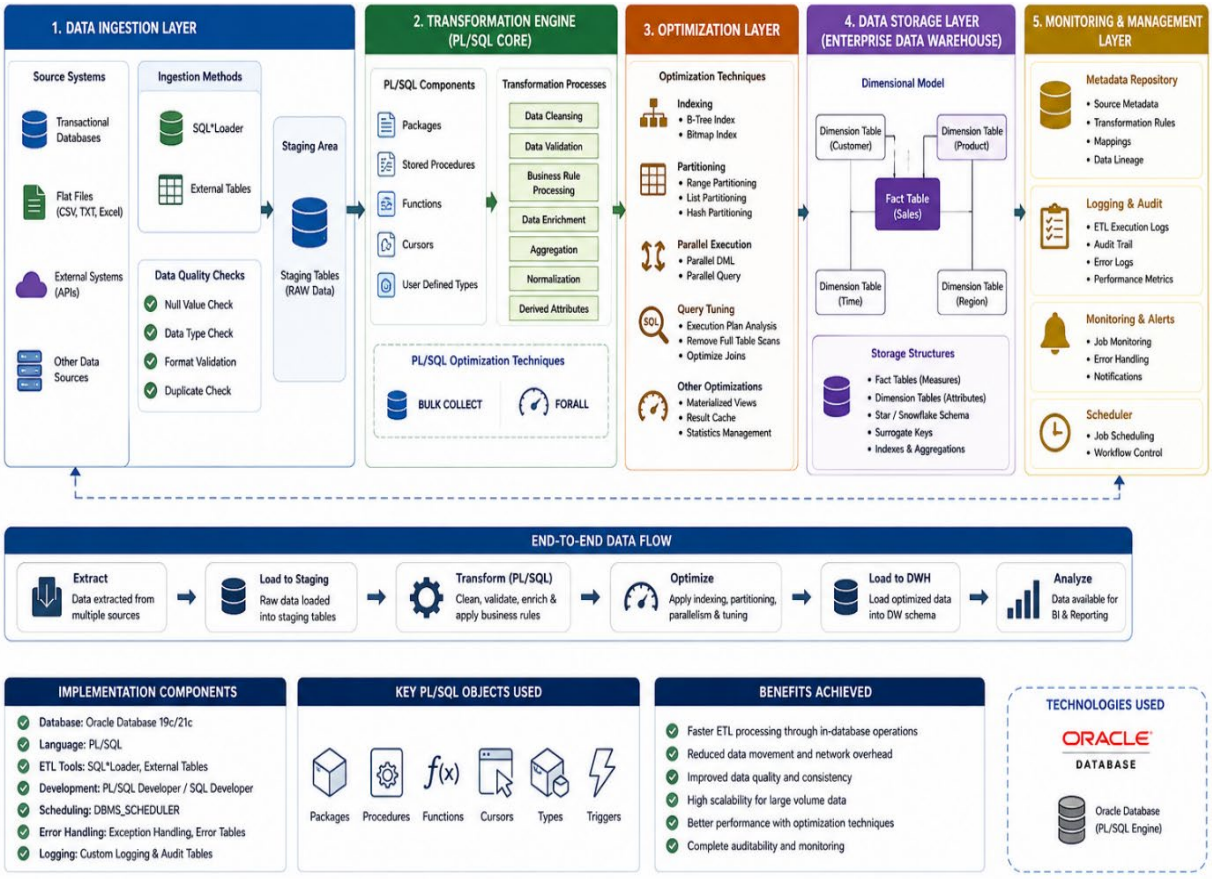


Figure 2. Implementation of a PL/SQL-Based Data Transformation and Optimization Framework for Enterprise Data Warehouses

The PL/SQL-based Data Transformation and Optimization Framework for enterprise data warehouses is implemented in phases using Oracle Database. In the first stage, the data is extracted from several enterprise sources, including transactional databases, flat files, and external systems. Raw data is imported into the staging

tables using Oracle utilities such as SQL*Loader and External Tables. During this stage, data validation checks are performed to identify null records, duplicates, bad data formats, and wrong data types. In the second phase, the data is processed in the staging layer, where it is stored in temporary tables. Temporary tables play a vital role in separating source data from warehouse tables and in allowing for the review of failed or rejected records. Implementation at the Core Level is facilitated by the PL/SQL Transformation Engine that consists of PL/SQL packages, procedures, functions, and cursors. The engine facilitates data cleansing, validation, standardization, enrichment, aggregation, normalization, and business rule enforcement. Techniques such as BULK COLLECT and FORALL facilitate processing voluminous data by reducing context switching between the SQL and PL/SQL engines. The third phase focuses on performance enhancement through indexing, partitioning, parallel processing, and query tuning. For instance, B-tree indexes are implemented on high-cardinality columns, whereas bitmap indexes are deployed on low-cardinality columns. Tables containing millions of records are partitioned to enhance query and transformation performance. Finally, the transformation and optimization processes facilitate loading dimensional models into the data warehouse. In this phase, fact tables will contain measurable business data, while dimensional tables will contain descriptive data. Log files, audit tables, metadata tracking, and error-handling mechanisms will facilitate execution monitoring, failure capture, and record reprocessing.

VI.RESULTS AND EVALUATION

The proposed framework was tested using a real-world banking data warehouse environment, specifically a core banking data warehouse. The data warehouse application is designed to handle large volumes of transaction data, including account, credit card, loan, and ATM transactions.

1. Case Study of Real-Time Banking

The case study involves a large bank that processes 10-15 million transactions daily across various channels, including mobile and internet banking, ATMs, and branch operations. Under the ETL methodology, the data was extracted, transformed using external ETL software, and subsequently loaded into the warehouse. The framework supports real-time banking and other applications where large volumes of data, minimal latency, regulatory requirements, and real-time analysis are key requirements.

2. Experimental Setup

The experiments were performed using Oracle Database 19c (19.11) on a dedicated Exadata X8M infrastructure with identical hardware configurations: 2-node RAC clusters with 48-core Intel Xeon processors, 768 GB of RAM per node, and high-performance NVMe flash storage. The benchmark was designed using a star schema model that included one transactional fact table and several dimension tables for customers, accounts, branches, products, and time. The workload included banking transactions, such as account transfers, card transactions, loan payments, ATM withdrawals, and branch transactions. All experiments were conducted five times, and the average result was calculated. The baseline approaches include an ETL process in an external environment, row-by-row processing using PL/SQL, and set-based transformation of SQL queries. The proposed approach is PL/SQL packages with BULK COLLECT, FORALL, partition pruning, indexing, and parallelization.

3. Results

As shown in the evaluation results, the new framework improves ETL performance compared with the legacy baseline method and the optimized PL/SQL approach. The average ETL execution time decreased from 8.4

hours to 3.1 hours, and the average query response time decreased from 47.2 seconds to 17.4 seconds, resulting in a 63.1% improvement. Peak ETL throughput increased from 285,000 rows per second to 1,840,000 rows per second, which is a 6.5-fold improvement. Concurrently, the data load error rate decreased from 3.8% to 2.0%. Finally, the PGA memory waste decreased from 34.2% to 8.1%. Additionally, the framework improved data quality through integrated error handling, logging, and audit mechanisms.

Metric	Legacy Baseline	Bulk-Optimized PL/SQL	NPDTOF	Improvement vs. Baseline
Avg. ETL Elapsed Time (hours)	8.4	5.9	3.1	63.1% reduction
Avg. Query Response Time (seconds)	47.2	31.6	17.4	63.1% reduction
Peak Throughput (rows/sec)	285,000	890,000	1,840,000	6.5x increase
Data Load Error Rate (%)	3.8%	2.1%	2.0%	47.4% reduction
PGA Memory Waste (%)	34.2%	18.7%	8.1%	76.3% reduction

Figure 3. Performance Comparison Across Evaluation Environments (Averaged)

4. Discussion

The experimental outcomes demonstrate considerable and multifaceted advantages over conventional PL/SQL ETL solutions. It should be noted that the level of improvement achieved – especially 6.5 times faster ETL processes and 63% query time reduction – is not provided by any specific approach; rather, the efficiency of this framework is attributable to the integrated use of bulk processing, flexible memory management, partition-aware processing, and metadata-based approach within the framework of transformations.

Another noteworthy finding is the substantial performance improvement observed in the banking workload. This environment showed the largest query time reduction, at 71%, despite having the lowest data volume among the testing environments. The analysis indicates that the banking data transformation workload involves very complex tasks of lookup, validation, and referencing due to customer profiles, account hierarchy, bank branch maps, product codes, transaction classification, and regulation- and risk-related reference tables. The integration of caching and dimension pre-loading proved particularly effective in reducing redundant lookup operations and queries and enabling fast transformation of account-, transaction-, customer-, and compliance-related data.

VII.LIMITATIONS

The use of Oracle and PL/SQL in the proposed framework results in limited portability, as it cannot be ported to another platform. This method is primarily focused on batch processing; therefore, integrating data in real time or via streams is difficult. The suggested solution requires knowledge of PL/SQL, making its implementation and maintenance technically complex. Even though the proposed framework improves performance when implemented in a single database, it lacks scalability when multiple databases are involved. Using database resources may limit performance if the hardware is underpowered.

VIII.FUTURE WORK

Future research should implement real-time and streaming capabilities, reduce Oracle dependency by ensuring cloud computing and platform independence, and increase scalability by integrating big data platforms such as Spark. Moreover, future research aims to incorporate AI-based optimization, improve data quality and governance, design user-friendly interfaces, and enhance security capabilities.

IX.CONCLUSION

This paper describes a new PL/SQL-based data transformation and optimization framework for enterprise data warehouses, designed to improve ETL efficiency through in-database processing. By leveraging Oracle's capabilities, such as PL/SQL procedures, BULK COLLECT, FORALL, indexing, partitioning, and parallelism, the proposed framework reduces data movement and improves performance. The multi-layered framework, comprising data ingestion, transformation, optimization, and storage layers, ensures systematic and efficient data processing. Experiments conducted in this paper and case studies have shown measurable improvements in processing time, resource utilization, data quality, and fault tolerance. In summary, this study offers effective solutions for high-volume Oracle-based workloads within enterprises. Even though the framework has some limitations regarding its reliance on a particular platform and its real-time capabilities, it provides a robust foundation for efficient ETL processing.

REFERENCES

[1] Amazon Web Services. (2024). What is ETL? https://aws.amazon.com/what-is/etl/
[2] Google Cloud. (2024). What is ETL? https://cloud.google.com/learn/what-is-etl
[3] Microsoft. (2024). ETL vs ELT. https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl
[4] Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data. Wiley. https://www.wiley.com/en-us/The+Data+Warehouse+ETL+Toolkit-p-9780764567575
[5] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual modeling for ETL processes," in Proc. DOLAP, 2002, pp. 14–21.
[6] X. Liu, "Optimizing ETL dataflow using shared caching and parallelization methods," arXiv preprint arXiv:1409.1639, 2014.
[7] A. Simitsis, P. Vassiliadis, and T. Sellis, "Optimizing ETL processes in data warehouses," in Proc. VLDB Endowment, 2005.
[8] R. K. Kodali et al., "Push down optimization for distributed multi-cloud data integration," arXiv preprint arXiv:2601.17546, 2026.
[9] Yu, X., Youill, M., Woicik, M., Ghanem, A., Serafini, M., Aboulnaga, A., & Stonebraker, M. (2020). PushdownDB: Accelerating a DBMS using S3 computation. IEEE ICDE 2020, 1802–1805. DOI: 10.1109/ICDE48307.2020.00174.
[10] R. Zhang et al., "Optimal predicate pushdown synthesis," arXiv preprint arXiv:2604.13351, 2026.
[11] N. Ebadifard et al., "Data extraction, transformation, and loading process automation for machine learning modeling," arXiv preprint arXiv:2312.12774, 2023.
[12] IBM, "ETL (Extract, Transform, Load)," 2024. [Online]. Available: https://www.ibm.com/think/topics/etl
[13] Feuerstein, S., & Pribyl, B. (2014). Oracle PL/SQL programming (6th ed.). O'Reilly Media.
[14] Harrison, G. (2010). Oracle Performance Survival Guide. Prentice Hall.
[15] Dinesh, L., & Devi, K. G. (2024). An efficient hybrid optimization of the ETL process in a cloud data warehouse. Journal of Cloud Computing, 13, Article 12. DOI: 10.1186/s13677-023-00571-y.



[16] Khan, B., Jan, S., Khan, W., & Chughtai, M. I. (2024). An overview of ETL techniques, tools, processes, and evaluations in data warehousing. Journal on Big Data. 6(1), 1–20. DOI: 10.32604/jbd. 2023.046223.
[17] Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 207, 111855. DOI: 10.1016/j.jss.2023.111855.
[18] Yin, S., Morvan, F., Martinez-Gil, J., & Hameurlain, A. (2025). MTD-DS: An SLA-aware decision support benchmark for multi-tenant parallel DBMSs. IEEE Transactions on Knowledge and Data Engineering, 37(5), 2743–2755. DOI: 10.1109/TKDE.2025.3543727.
[19] Pasupuleti, K. K., Li, J., Su, H., & Ziauddin, M. (2023). Automatic SQL error mitigation in Oracle. Proceedings of the VLDB Endowment, 16(12), 3835–3847. DOI: 10.14778/3611540.3611568.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Multi-Agent Conflict Simulation With World Constraints (MACSWC)

Borjan Ladinski
Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
borjan.ladinski@gmail.com
ORCID: 0009-0004-9916-4618

Dusko Sofronievski
Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
dukisofronievski@gmail.com
ORCID: 0009-0002-6790-0225

Marija Stankova Medarovska
Computer Science and Information Technology
University American College Skopje
Skopje, North Macedonia
marija.s.medarovska@uacs.edu.mk
ORCID: 0009-0000-0875-7911

Abstract
Large Language Models (LLMs) have enabled the development of intelligent multi-agent systems capable of planning, collaborating, and making strategic decisions. Recent frameworks have demonstrated the feasibility of employing LLMs as autonomous agents in many domains. However, directly allowing language models to modify persistent environments introduces challenges, including hallucinated actions, invalid state transitions, and inconsistent world representations. These issues become particularly critical in geopolitical simulations, where every decision permanently affects future interactions. This paper presents MACSWC (Multi-Agent Conflict Simulation with World Constraints), a deterministic multi-agent framework designed for safe geopolitical simulation. Inspired by the Second Balkan War, it models sovereign-state agents that make diplomatic, military, and economic decisions using LLMs and operate under a layered validation architecture. Unlike existing systems that directly execute model outputs, MACSWC treats every LLM-generated action as a proposal requiring deterministic validation before execution. The framework consists of Country Agents for strategic reasoning, a Secretary Agent for structural validation, a Board Agent for strategic evaluation, and a centralized World Engine that maintains the authoritative simulation state. This separation between reasoning and execution improves consistency, prevents invalid state transitions, and enables reproducible experimentation across multiple language model providers. The proposed architecture contributes to a reusable framework for trustworthy autonomous simulations and demonstrates how deterministic governance can improve the reliability of LLM-based multi-agent systems.

Keywords: Large Language Models, Multi-Agent Systems, Autonomous Agents, Geopolitical Simulation, Deterministic Validation, AI Governance

I. INTRODUCTION

The rapid development of Large Language Models (LLMs) has transformed artificial intelligence by enabling systems capable of sophisticated reasoning, planning, and natural language understanding. Foundation models such as GPT-4 have demonstrated performance across a wide range of cognitive tasks, encouraging researchers to investigate their use as autonomous decision-making agents rather than conversational assistants alone [1], [2]. One of the fastest-growing research areas is Large Language Model Multi-Agent Systems (LLM-MAS), in which multiple autonomous agents cooperate or compete to solve complex problems. Frameworks such as CAMEL [3], AutoGen [4], MetaGPT [5], and Generative Agents [6] have shown that specialized LLM agents can collaborate effectively through role-based communication, allowing them to complete tasks beyond the capabilities of individual models. More recently, WarAgent demonstrated that autonomous LLM agents could simulate historical geopolitical conflicts through diplomatic negotiation and strategic decision-making [7].

Despite these advances, existing LLM-based systems remain fundamentally probabilistic. Language models may generate hallucinated information, malformed structured outputs, contradictory reasoning, or actions that violate environmental constraints [8]. While such errors may be acceptable in conversational applications, they become

problematic in persistent simulations where every decision modifies a shared world state. A single invalid action may introduce inconsistencies that propagate throughout the remainder of the simulation. Several approaches have attempted to improve reasoning quality through prompt engineering, role specialization, reflection, or self-correction [9-11]. Although these methods reduce reasoning errors, they do not guarantee that the generated actions satisfy the environment's logical or structural requirements. Consequently, the integrity of the simulation often remains dependent on the correctness of language model outputs. This paper introduces MACSWC (Multi-Agent Conflict Simulation with World Constraints), a deterministic framework that separates strategic reasoning from authoritative world-state management. Rather than allowing LLMs to modify the environment directly, MACSWC introduces multiple validation layers that evaluate every proposed action before execution. Country Agents generate strategic proposals, a Secretary Agent validates their correctness, a Board Agent evaluates strategic utility and risk, and a centralized World Engine performs all state modifications. The simulation environment is inspired by the Second Balkan War of 1913 and models interactions among Bulgaria, Serbia, Greece, Romania, and the Ottoman Empire. Although the historical setting provides a realistic geopolitical scenario, the primary contribution of this work is not historical reconstruction but the introduction of deterministic governance for autonomous LLM-based simulations. The contributions of this work are summarized as follows:

- A deterministic architecture that separates LLM reasoning from world-state authority.
- Layered validation pipeline consisting of Country Agents, a Secretary Agent, a Board Agent, and a World Engine.
- Configurable geopolitical simulation supporting diplomatic, military, and economic interactions.
- Model-independent framework compatible with multiple LLM providers.
- Reusable architecture for trustworthy multi-agent simulations.

The remainder of this paper is organized as follows. Section II reviews related research on LLM-based multi-agent systems and AI safety. Section III presents the MACSWC framework. Section IV describes the experimental evaluation, followed by the obtained results and conclusions.

II. RELATED WORK

Recent advances in LLM-based multi-agent systems have demonstrated that language models can cooperate to solve complex reasoning tasks. CAMEL introduced autonomous role-playing agents that collaborate through structured prompting [3], while AutoGen proposed conversational agents capable of iterative problem solving [4]. MetaGPT extended this concept by organizing agents into software engineering roles, demonstrating that role specialization improves task coordination [5]. Similarly, Generative Agents showed that LLMs can maintain memories, plan future activities, and interact realistically within persistent virtual environments [6]. Beyond collaborative reasoning, researchers have investigated the use of LLMs for strategic simulation. WarAgent is among the first large-scale attempts to simulate historical geopolitical conflicts using autonomous country agents [7]. The framework demonstrated that language models can produce coherent diplomatic negotiations and military strategies, enabling the exploration of alternative historical outcomes. However, the simulation's correctness remains strongly dependent on the outputs of the underlying language models. Ensuring the reliable behavior of autonomous language models has therefore become an important research topic. Constitutional AI [9], Reinforcement Learning from Human Feedback (RLHF) [10], ReAct [11], Reflexion [12], and Tree of Thoughts [13] improve reasoning quality through feedback, self-reflection, or structured planning. Although these approaches significantly reduce reasoning errors, they focus primarily on improving the language model itself rather than protecting the environment from incorrect outputs. The key distinction of MACSWC is the introduction of deterministic world-state governance. Rather than assuming that language model outputs are correct, every generated action is treated as a proposal that requires independent validation before execution.

This design separates probabilistic reasoning from authoritative state management, improving reproducibility while preventing hallucinated or logically inconsistent actions from modifying the simulation.

III. MACSWC FRAMEWORK

1. System Architecture

MACSWC is designed around the principle that Large Language Models perform reasoning but never directly modify the simulation state. Instead, every action generated by an LLM is interpreted as a proposal that must pass multiple deterministic validation stages before execution. This separation between reasoning and authority improves reliability while preserving the flexibility of autonomous decision-making. The framework consists of four primary components: Country Agents, a Secretary Agent, a Board Agent, and a centralized World Engine. Together, these components form a layered decision pipeline that validates every action before it affects the simulated environment. During each simulation round, Country Agents receive a structured observation describing the current geopolitical situation. Based on this information, each agent generates one strategic proposal. The proposal is then validated by the Secretary Agent, evaluated by the Board Agent, and, if all validation stages are completed, executed by the World Engine. Fig.1 illustrates the complete architecture.

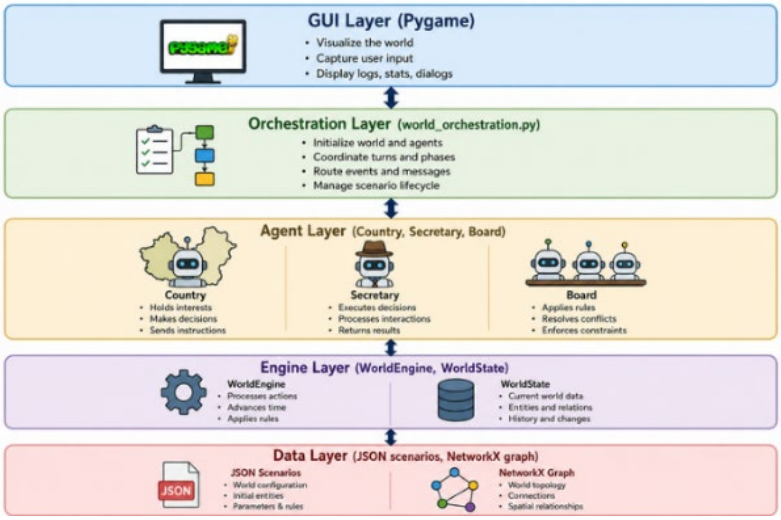


Figure 1. System Architecture.

2. Country Agents

Country Agents represent sovereign states participating in the simulation. Each agent operates independently and receives a structured observation containing diplomatic relations, military status, alliances, economic conditions, recent historical events, and currently available actions. Using this information, the agent selects one diplomatic, military, or economic action for the current simulation round. Supported actions include alliance formation, declarations of war, peace negotiations, sanctions, trade agreements, and diplomatic communication. MACSWC is model-independent and supports multiple LLM providers, including OpenAI, Groq, Hugging

Face, and Ollama. Since every model communicates through the same interface, different LLMs can be evaluated under identical simulation conditions.

3. Secretary Agent

The Secretary Agent performs deterministic validation before strategic evaluation. Its responsibilities include verifying the JSON structure, required parameters, the legality of actions, target countries, and compliance with the registered action set. Invalid proposals are rejected immediately, preventing malformed outputs or hallucinated entities from entering the simulation. This validation layer isolates common LLM failures from the remaining components and ensures that only structurally valid proposals proceed to strategic evaluation.

4. Board Agent

The Board Agent evaluates proposals that satisfy structural validation. The Board Agent estimates the strategic value of each proposal by considering diplomatic relationships, alliance structures, military conflicts, and crisis risk. A simplified utility function is used to determine whether the expected benefit outweighs the potential geopolitical cost:

$$U = W - \alpha R$$

where W denotes the estimated strategic benefit, R represents crisis risk, and α controls the influence of risk on the final utility score. Only proposals exceeding the approval threshold are forwarded to the execution layer.

5. World Engine

The World Engine represents the simulation's authoritative state and is the only component permitted to modify the environment. It maintains all countries, alliances, diplomatic relations, military conflicts, sanctions, treaties, and historical events. Before executing an approved proposal, the World Engine performs a final consistency check to ensure that the resulting state transition remains compatible with the current geopolitical situation. Centralizing all modifications prevents contradictory state transitions and guarantees that every Country Agent observes a consistent representation of the simulated world.

6. Decision Pipeline

Each simulation round follows the same deterministic workflow:

- The World Engine generates structured observations.
- Country Agents generate strategic proposals.
- The Secretary Agent validates structural correctness.
- The Board Agent evaluates strategic utility.
- The World Engine executes approved actions.
- The updated world state becomes the input for the following round.

Sequential execution ensures deterministic behavior while preventing conflicting updates that could occur during simultaneous execution.

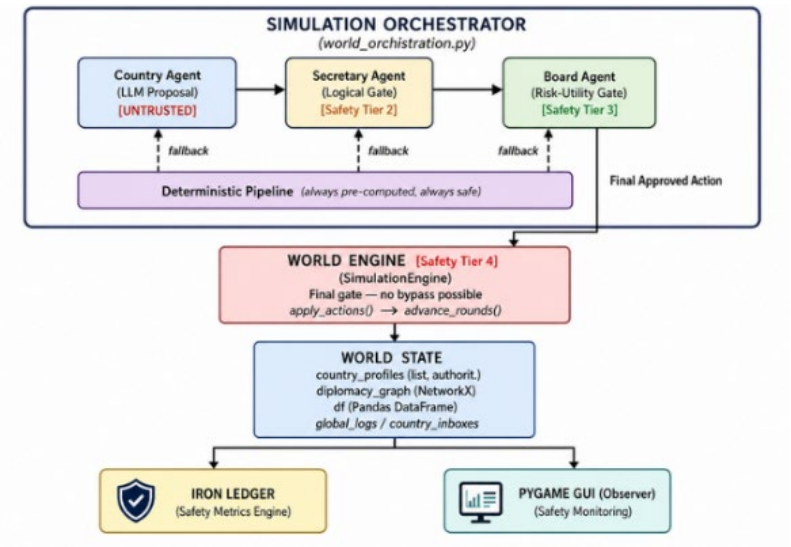


Figure 2. Layered software organization of MACSWC, illustrating the separation between the data, execution, agent, orchestration, and presentation layers to improve modularity and simplify future extension.

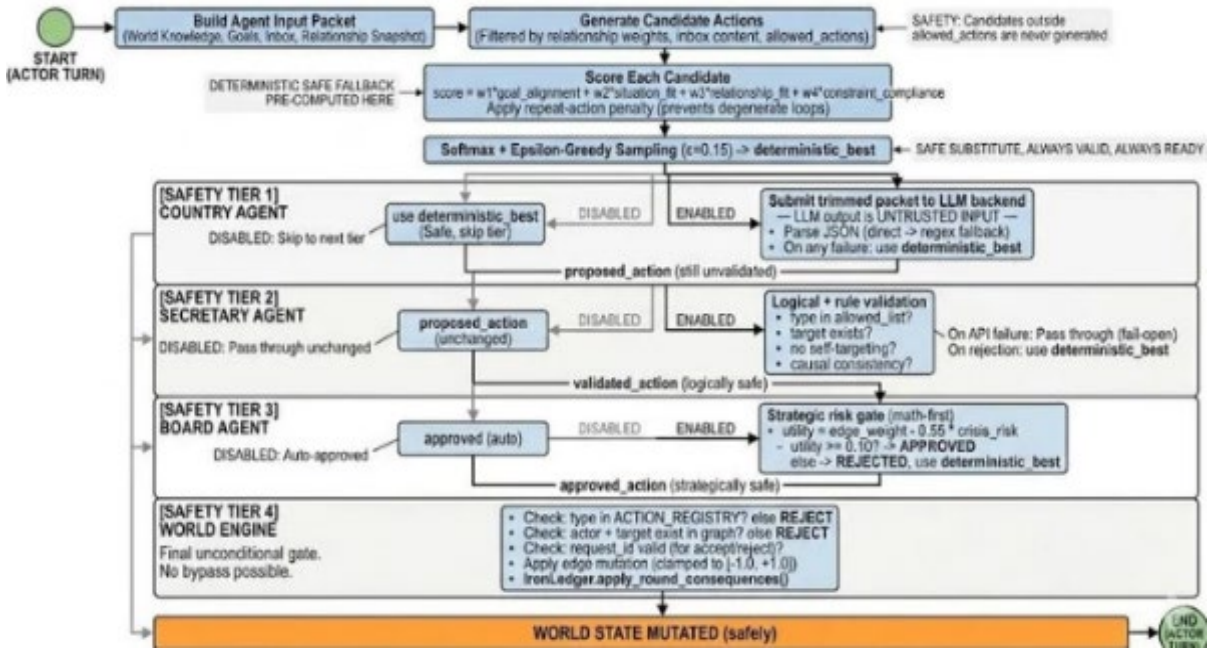


Figure 3. Deterministic decision pipeline illustrating proposal generation, validation, strategic evaluation, and authoritative world-state execution.

IV. EXPERIMENTAL SETUP

The MACSWC framework was implemented in Python 3 using a modular architecture. NetworkX was used to represent diplomatic relationship graphs, while JSON files maintained persistent world-state information. The modular implementation allows individual components to be replaced without modifying the remaining architecture. The evaluation scenario is inspired by the Second Balkan War of 1913 and includes five autonomous Country Agents representing Bulgaria, Serbia, Greece, Romania, and the Ottoman Empire. Each country independently generates strategic actions throughout multiple simulation rounds. The framework supports several LLM providers through a unified interface, including OpenAI GPT models, Groq-hosted models, Hugging Face inference endpoints, and locally deployed Ollama models. Since all models interact with the same deterministic validation pipeline, the architecture remains independent of the underlying reasoning model. The evaluation focuses on architectural reliability rather than prediction accuracy. During execution, the framework records accepted proposals, rejected proposals, alliance formation, military conflicts, diplomatic changes, crisis risk, and simulation stability. These metrics enable reproducible comparison between different simulation runs and language models.



Figure 4. Graphical user interface of MACSWC displaying the geopolitical map, diplomatic relationship graph, country status, and event log during a simulation round.

V. RESULTS AND DISCUSSION

The experimental evaluation demonstrates that deterministic governance substantially improves the reliability of autonomous LLM-based geopolitical simulations. The Secretary Agent successfully intercepted malformed outputs, unsupported actions, missing parameters, and hallucinated entities before they reached the execution layer. Consequently, structural errors generated by language models did not affect the simulation state. The Board Agent provided an additional safety mechanism by rejecting proposals associated with excessive geopolitical risk. Although structurally valid, several actions were prevented from executing because their estimated strategic cost exceeded the predefined approval threshold. This reduced unnecessary escalation while preserving adaptive strategic behavior. The World Engine maintained a consistent and reproducible simulation throughout all experimental rounds. Every accepted proposal satisfied structural, logical, and environmental

constraints before execution, preventing contradictory diplomatic relationships, duplicate alliances, and impossible military operations. Compared with recent LLM-based multi-agent frameworks such as CAMEL [3], AutoGen [4], MetaGPT [5], and WarAgent [7], MACSWC contributes deterministic world-state governance rather than improvements in reasoning capability. While existing systems primarily investigate autonomous collaboration, MACSWC demonstrates that separating reasoning from execution significantly improves simulation consistency without restricting the flexibility of language models. The modular architecture further enables future expansion through additional countries, alternative historical scenarios, richer economic systems, or different language model providers without requiring changes to the validation pipeline. Consequently, MACSWC serves not only as a geopolitical simulator but also as a reusable platform for trustworthy autonomous multi-agent research.

Table 1. Simulation Performance Metrics

Metric	**Value**
Simulation rounds	50
Actions generated	250
Accepted actions	201
Secretary rejections	31
Board rejections	18
World Engine rejections	0

VI. CONCLUSION

This paper presented MACSWC, a deterministic multi-agent framework designed to integrate Large Language Models into persistent geopolitical simulations safely. Unlike conventional LLM-based systems that directly execute model outputs, MACSWC separates strategic reasoning from authoritative execution via a layered architecture comprising Country Agents, a Secretary Agent, a Board Agent, and a centralized World Engine. Experimental evaluation demonstrated that deterministic validation improves simulation reliability by preventing malformed actions, logical inconsistencies, and invalid world-state transitions. The proposed architecture remains compatible with multiple LLM providers while maintaining a consistent and reproducible simulation environment. The principal contribution of MACSWC is the introduction of deterministic world-state governance for autonomous language model agents. By treating every generated action as a proposal that requires independent validation rather than as an executable command, the framework combines the reasoning capabilities of modern LLMs with the reliability of deterministic software systems. Future work will investigate broader geopolitical scenarios, adaptive strategic evaluation, and human-in-the-loop decision support to extend the framework's applicability further.

REFERENCES

[1] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.
[2] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.
[3] G. Li, H. Hammoud, H. Itani, B. Khizbullin, and B. Ghanem, “CAMEL: Communicative Agents for ‘Mind’ Exploration of Large Language Model Society,” Advances in Neural Information Processing Systems Workshops, 2023.

[4] Q. Wu et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," International Conference on Learning Representations (ICLR) Workshop, 2024.
[5] S. Hong et al., "MetaGPT: Meta Programming for Multi-Agent Collaborative Framework," International Conference on Learning Representations (ICLR), 2024.
[6] J. Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," Proceedings of the ACM Symposium on User Interface Software and Technology (UIST), pp. 1–22, 2023.
[7] W. Hua et al., "WarAgent: Large Language Model-Based Multi-Agent Simulation of World Wars," arXiv preprint arXiv:2311.17227, 2024.
[8] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, 2023.
[9] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv preprint arXiv:2212.08073, 2022.
[10] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.
[11] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," International Conference on Learning Representations (ICLR), 2023.
[12] N. Shinn et al., "Reflexion: Language Agents with Verbal Reinforcement Learning," Advances in Neural Information Processing Systems (NeurIPS), 2023.
[13] S. Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," Advances in Neural Information Processing Systems (NeurIPS), 2023.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Trustworthiness Under Compression: Calibration and Explanation Faithfulness of Lightweight IoT Intrusion Detectors

Md Anas Biswas
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
up2082724@myport.ac.uk
ORCID: 0009-0009-0113-5816

Alif Laila
School of Computing
University of Portsmouth
Portsmouth, United Kingdom
up2607085@myport.ac.uk
ORCID: 0009-0004-9633-8969

Abstract
Compressed Internet-of-Things intrusion detectors are usually evaluated by aggregate accuracy, which can mask failures during rare attacks. We audit a compact multilayer perceptron and quantized, pruned, and distilled variants on the full 46-million-flow CICIoT2023 dataset and TON_IoT. Per-class calibration is measured using adaptive expected calibration error and signed confidence gaps; explanation agreement is assessed using KernelSHAP against repeatability and retraining references, as well as through feature-deletion tests. Aggregate accuracy remains high, yet aggressive pruning causes rare-class overconfidence, attribution drift, and zero recall for Web and BruteForce attacks, which it predicts as benign. The attributions remain behaviourally faithful, indicating a change in feature reliance rather than an explainer failure. The pattern transfers across datasets and an independent split. Temperature scaling reduces calibration error but cannot recover lost recall. Compressed detectors should therefore be accepted only after per-class calibration, recall, and explanation-faithfulness checks.

Keywords: class imbalance, long-tail evaluation, intrusion detection, model compression, calibration, explainable AI, trustworthy machine learning.

I. INTRODUCTION

Modern IoT intrusion benchmarks are large and long-tailed. CICIoT2023 contains roughly 46 million flows [1], yet a few classes dominate while security-critical attacks occupy the tail. Aggregate accuracy and weighted F1 are therefore weak acceptance criteria: a detector can score well while failing on rare classes. This work tests whether compression conceals that failure. IoT intrusion detection systems (IDSs) are deployed on gateways with strict memory and compute budgets. Quantization, pruning, and knowledge distillation reduce deployment cost, but compressed models are commonly evaluated only by retained accuracy. Trust also requires calibrated confidence and faithful explanations: alerts should be probabilistically credible and traceable to the computation that produced them. We audit a compact multilayer perceptron (MLP) and five compressed variants on CICIoT2023 and TON_IoT [2]. Per-class calibration is paired with SHAP attribution agreement and deletion tests. Aggressive pruning preserves aggregate accuracy while rare attacks become over-confident, explanations drift, and some attacks are reassigned to benign. Temperature scaling improves calibration but not recall. Our contributions are: (1) a per-class audit that exposes long-tail failures hidden by aggregate metrics; (2) a paired calibration and explanation-faithfulness protocol with bootstrap intervals and repeatability/retraining references; and (3) evidence across two datasets and an independent split, including a benign blind spot, a feature-reliance shift, and a calibration-only remedy.

II. RELATED WORK

1. Lightweight and compressed intrusion detection

Model-compression techniques used for edge deployment include quantization [3], knowledge distillation [4], and pruning [5]. These foundational studies report favorable accuracy–cost trade-offs in general machine-learning settings; evidence specific to compressed IoT intrusion detectors remains limited. Pandey and Ye [6] further show that 8-bit integer (INT8) quantization can turn off backdoor defenses, even while attacks remain effective. We instead examine per-class calibration and explanation faithfulness of the deployed detector.

2. Calibration of neural networks and security models

Neural networks can be overconfident [7], and temperature scaling is a simple post hoc remedy. In imbalanced security data, aggregate calibration can conceal confident errors on rare attacks. We therefore report adaptive expected calibration error and a signed per-class confidence gap.

3. Explanation of faithfulness and stability

SHAP (SHapley Additive exPlanations) [8] is widely used for intrusion-detector interpretation. Security studies evaluate the quality of explanations [9]; Kalakoti et al. [10] compare SHAP, Local Interpretable Model-agnostic Explanations (LIME), Integrated Gradients [11], and DeepLIFT for network intrusion alerts. Interpretations can be fragile [12]. Adversarial behavior can be implemented using the Adversarial Robustness Toolbox (ART) [13], demonstrated with one-pixel perturbations [14], and it affects intrusion detectors [15]; Rajhans and Khawarey [16] directly analyze cybersecurity explanation drift. We study compression, compare against repeatability/retraining references, and verify attribution faithfulness through feature deletion.

III. METHODOLOGY

1. Datasets and labels

CICIoT2023 [1] is the primary benchmark: approximately 46 million flows, 33 attack types mapped to seven groups plus benign, and 46 features. In its test partition, distributed denial-of-service (DDoS) exceeds 5 million flows, whereas Web and BruteForce contain 3,724 and 1,960, respectively. The TON_IoT network subset [2] contains 461,043 flows, a normal class plus nine attacks, and 74 post-processed features. We remove identity-revealing fields, coerce missing markers, one-hot encode low-cardinality categories, use locked stratified splits, fit scaling only on training data, and apply tempered square-root inverse-frequency class weights. No resampling is used, preserving calibration.

2. Model and compression matrix

The anchor is a compact MLP with 128- and 64-unit hidden layers, batch normalization, dropout, and 15,304 parameters. Seed zero is specified as the reference before trust analysis. Quantized variants use TensorFlow Lite (TFLite) dynamic-range INT8 and float16; pruned variants use fixed 50% and 80% magnitude masks with fine-tuning. The distilled student has 3,632 parameters and combines softened teacher logits with hard labels. Quantization and pruning retain the anchor architecture; distillation changes the student architecture and is therefore interpreted as a combined change in architecture and compression. Pruning is evaluated as sparsification and does not reduce dense file size.

3. Calibration measurement

Calibration is measured with adaptive, equal-mass expected calibration error (ECE) at 10, 15, and 20 bins and a signed gap, mean confidence minus accuracy, where positive values indicate over-confidence. Metrics are reported per class, with bootstrap 95% confidence intervals for the signed gaps. Point estimates use all observations. A single temperature is fitted on validation data and evaluated on the locked test set.

4. Explanation of the faithfulness measurement

KernelSHAP is applied via a single probability interface to every model, using 80 seeded instances per class, a fixed k-means background, and 1,500 samples. Attributions target each instance's true class. Agreement is the per-instance Spearman correlation with the anchor, interpreted against a same-model repeatability reference and independently seeded retraining reference. Drift is declared only when its bootstrap interval lies below the retraining reference. Faithfulness is tested by deleting top-ranked features across three maskers and comparing the drop in true-class probability with that from random deletion. Web and BruteForce estimates are interpreted with caution because each uses 80 explained instances, and baseline recall is low.

5. Reproducibility

All splits and stochastic stages use fixed seeds. The scaler is fitted on training data only, the anchor is fixed before the trust analysis, and identical calibration and explanation protocols are applied across variants and datasets. Analysis artifacts and environment specifications are retained to support regeneration of the reported results.

IV. RESULTS

1. Accuracy and aggregate calibration are retained

Table 1 shows that accuracy stays near 0.99 except for INT8 (0.937), including 0.978 under 80% pruning. Macro F1 falls from 0.772 to 0.541, exposing rare-class loss. Only quantization and distillation reduce file size; pruning leaves dense storage unchanged.

Table 1. Compression outcomes: pruning zero weights without reducing the dense file size.

Model	Size (MB)	Macro-F1	Accuracy
Baseline (M0)	0.217	0.772	0.992
50% pruning	0.217*	0.720	0.991
80% pruning	0.217*	0.541	0.978
Distilled student	0.039	0.703	0.990
INT8 quantisation	0.021	0.691	0.937
Float16 quantisation	0.033	0.772	0.992

Aggregate ECE is also reassuring, increasing only from 0.003 to 0.014 under 80% pruning. The per-class results below show why this aggregate is insufficient.

2. Per-class calibration collapses on the rare classes

Table 2 shows the signed confidence gap. Under 80% pruning, Web rises from 0.044 to 0.575, BruteForce from 0.171 to 0.600, Spoofing from 0.007 to 0.412, and Recon from 0.032 to 0.395; DDoS, DoS, and Mirai change little. Figure 1 summarises aggregate ECE and attribution drift across variants relative to the retraining reference.

Table 2. Per-class signed confidence gap with bootstrap 95% confidence intervals (CIs) on CICIoT2023.

Class	Baseline gap [95% CI]	80% pruning gap [95% CI]
Benign	−0.120 [−0.122, −0.118]	−0.333 [−0.335, −0.333]
DDoS	−0.001 [−0.001, 0.000]	−0.010 [−0.010, −0.009]
DoS	0.000 [0.000, 0.001]	−0.041 [−0.042, −0.039]
Mirai	0.000 [−0.001, 0.000]	0.006 [0.005, 0.006]
Recon	0.032 [0.028, 0.035]	0.395 [0.390, 0.398]
Spoofing	0.007 [0.004, 0.010]	0.412 [0.410, 0.415]
Web	0.044 [0.030, 0.058]	0.575 [0.571, 0.578]
BruteForce	0.171 [0.154, 0.187]	0.600 [0.596, 0.604]

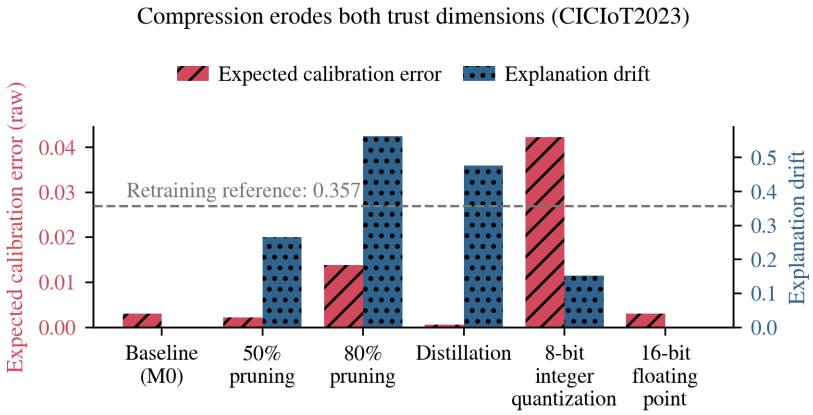


Figure 1. Aggregate expected calibration error and explanation drift by compression variant (CICIoT2023); M0 is the baseline, and the band is the retraining reference.

3. Explanations drift from the baseline but stay behaviourally faithful

Table 3 reports attribution agreement. The same-model repeatability reference is 0.773, and the retraining reference is 0.643. Agreement under 80% pruning (0.437) and distillation (0.524) lies below the retraining

reference, whereas float16 is effectively unchanged (0.999). In Figure 2, deleting top SHAP-ranked features produces a larger true-class probability drop than random deletion for both the anchor and pruned model. Compression changes feature reliance, but the resulting attributions remain behaviourally faithful.

Table 3. Attribution agreement with the baseline on CICIoT2023.

Comparison	Spearman [95% CI]	Below retraining reference
Same-model repeatability	0.773 [0.764, 0.782]	—
Retraining reference	0.643 [0.635, 0.651]	—
50% pruning	0.735 [0.723, 0.745]	No
80% pruning	0.437 [0.418, 0.455]	Yes
Distilled student	0.524 [0.510, 0.538]	Yes
INT8 quantisation	0.847 [0.833, 0.861]	No
Float16 quantisation	0.999 [0.998, 1.000]	No

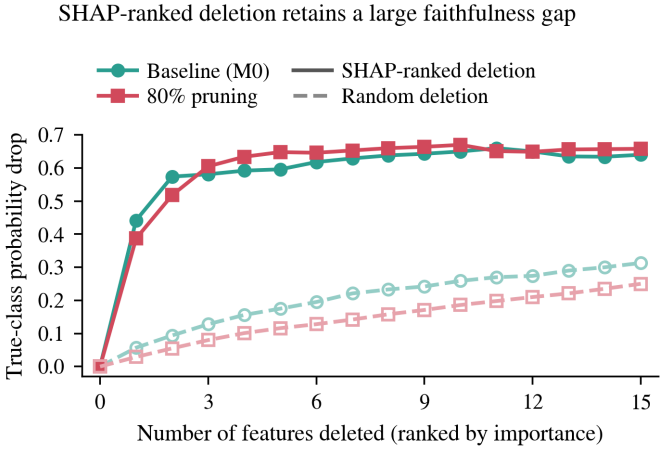


Figure 2. True-class probability drops after deleting SHAP-ranked or random features.

4. Calibration loss and explanation drift co-occur on the rare classes

Rare classes show both the largest confidence gaps and the lowest attribution agreement. Across eight classes, confidence gap and explanation drift correlate (Pearson r = 0.87, p = 0.005); after controlling for class size, the association is weaker (partial r = 0.64, p = 0.122). Within classes, per-instance over-confidence and drift are nearly unrelated (pooled Spearman rho = 0.04). The evidence, therefore, supports class-level co-occurrence through shared difficulty, not a common instance-level mechanism.

5. The benign blind spot

Table 4 and Figure 3 expose the most consequential failure. Baseline recall for Web and BruteForce is already limited (0.428 and 0.335), but 80% pruning reduces both to zero and predicts 95–96% of them as benign.

Spoofing and Recon also lose most recall to benign, while high-volume classes remain stable. Thus, a compressed detector can confidently miss rare attacks even as aggregate accuracy remains 0.978.

Table 4. Rare-class recall under 80% pruning on CICIoT2023.

Class	Baseline recall	80% pruning recall	Mostly predicted as
Web	0.428	0.000	Benign (0.95)
BruteForce	0.335	0.000	Benign (0.96)
Spoofing	0.769	0.147	Benign (0.85)
Recon	0.704	0.232	Benign (0.73)

Aggregate structure masks rare-class recall collapse under 80% pruning (CICIoT2023).

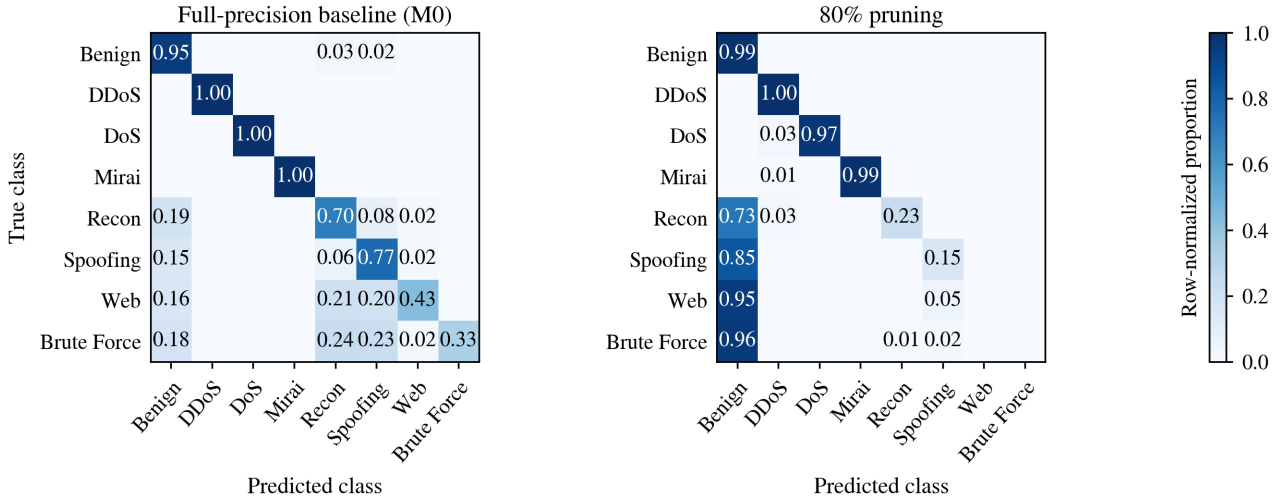


Figure 3. Row-normalized confusion matrices for the baseline and 80% pruning.

6. The pattern transfers to TON_IoT

The pattern transfers to TON_IoT (Figure 4). Aggregate ECE rises from 0.053 to 0.130 under 80% pruning; the largest confidence gap is Ransomware (0.668). Attribution agreement under pruning (0.451) and distillation (0.377) is below the 0.610 retraining reference, with the greatest drift on hard classes such as Injection, DoS, and Man-in-the-Middle. CICIoT2023 shows the same direction: ECE 0.003 to 0.014, worst gap 0.600, and pruning agreement 0.437 below its 0.643 reference.

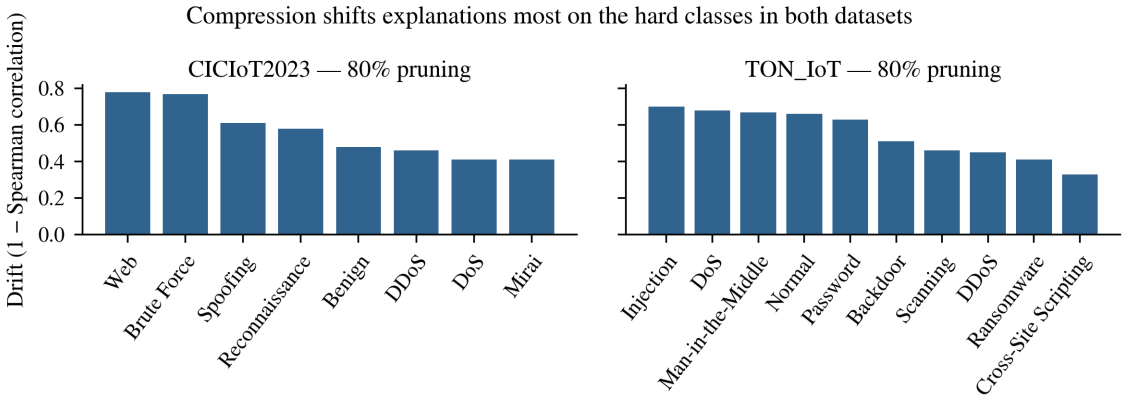


Figure 4. Per-class explanation drift under 80% pruning on CICIoT2023 and TON_IoT.

Dynamic-range INT8 is unstable on TON_IoT's 74-feature space, producing a degenerate fitted temperature; it is excluded from TON_IoT calibration reporting but retained in the attribution analysis. Pruning and distillation drive the cross-dataset conclusion.

7. Robustness and mechanism

An independent stratified split reproduces the pruning gaps (Web 0.585, BruteForce 0.589, Spoofing 0.558, and Recon 0.343), supporting split stability but not temporal generalization. Grouped SHAP importance also shifts modestly from packet-size statistics toward coarse flow-timing features under pruning, offering a candidate explanation for attribution drift.

8. A cheap remedy recovers calibration

Temperature scaling approximately halves CICIoT2023 ECE for 80% pruning (0.014 to 0.007) and INT8 (0.042 to 0.018), as Figure 5 shows. On TON_IoT, it helps every retained variant except distillation (0.063 to 0.084). Rescaling cannot restore rare-class recall because that loss reflects the decision boundary rather than confidence scaling.

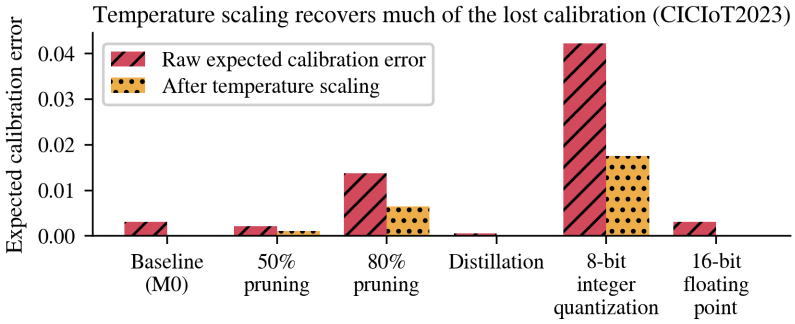


Figure 5. Expected calibration error before and after temperature scaling on CICIoT2023.

V. DISCUSSION

Aggregate accuracy is the most compression-resistant metric and, therefore, a poor proxy for security risk. Aggressive pruning produces over-confident rare-class failures, and shifts feature reliance; in the extreme, attacks are labeled benign and generate no alert. Per-class recall, calibration, and explanation behavior must therefore be reviewed together. Calibration loss and attribution drift co-occur on difficult classes but are not coupled within instances. Deletion tests show that explanations continue to track the compressed model, while the feature-family analysis suggests that pruning moves the model toward coarser timing signals. These findings support an operational association, not a single causal mechanism. Float16 is the safest tested option when trust is the priority: it is about seven times smaller with negligible trust loss. Distillation exhibits significant attribution drift and should be used only when a separately measured deployment benefit justifies its use. INT8 requires per-class calibration checks, and 80% pruning should be avoided where rare-class detection matters. Recommendations are based on file size and trust metrics; latency and energy were not measured.

VI. LIMITATIONS AND FUTURE WORK

The study fixes one MLP family to isolate the effects of quantization and pruning; the distilled student also changes its architecture. Split stability does not establish temporal or subtype-group generalization, and the TON_IoT INT8 calibration result is unstable. Model size is measured, but edge-device latency and energy are not. KernelSHAP uses 80 instances per class, so estimates for rare classes and any proposed feature-shift mechanism require further confirmation. Future work should test convolutional and recurrent detectors, full-integer per-channel quantization, temporal and subtype-grouped splits, larger explanation samples, and on-device latency and energy. A targeted causal study could determine whether calibration loss and attribution drift share a mechanism.

VII. CONCLUSION

Compressed IoT intrusion detectors can retain headline accuracy while becoming overconfident and misclassifying rare attacks as benign. Attribution drift remains behaviourally faithful, showing that compression changes the model's reliance rather than merely breaking the explainer. Temperature scaling reduces much of the confidence error, but not the recall error. Deployment decisions should therefore use per-class recall, calibration, and explanation-faithfulness checks against repeatability and retraining references, rather than aggregate accuracy alone.

REFERENCES

[1] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, “CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment,” Sensors, vol. 23, no. 13, Art. no. 5941, 2023, doi: 10.3390/s23135941.

[2] N. Moustafa, “A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets,” Sustainable Cities and Society, vol. 72, Art. no. 102994, 2021, doi: 10.1016/j.scs.2021.102994.

[3] B. Jacob et al., “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2704–2713.



[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[5] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. Int. Conf. Learning Representations (ICLR), 2016.

[6] R. Pandey and E. Ye, "Quantization blindspots: How model compression breaks backdoor defenses," arXiv:2512.06243, 2025, doi: 10.48550/arXiv. 2512.06243.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2017, pp. 1321–1330.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[9] A. Warnecke, D. Arp, C. Wressnegger, and K. Rieck, "Evaluating explanation methods for deep learning in security," in Proc. IEEE European Symp. Security and Privacy (EuroS&P), 2020, pp. 158–174.

[10] R. Kalakoti, R. Vaarandi, H. Bahşi, and S. Nõmm, "Evaluating explainable AI for deep learning-based network intrusion detection system alert classification," in Proc. 11th Int. Conf. Information Systems Security and Privacy (ICISSP), vol. 1, 2025, pp. 47–58, doi: 10.5220/0013180700003899.

[11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. Int. Conf. Machine Learning (ICML), 2017, pp. 3319–3328.

[12] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in Proc. ICML Workshop on Human Interpretability in Machine Learning, 2018.

[13] M.-I. Nicolae et al., "Adversarial robustness toolbox v1.0.0," arXiv preprint arXiv:1807.01069, 2018.

[14] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," IEEE Trans. Evol. Comput., vol. 23, no. 5, pp. 828–841, 2019.

[15] D. Han et al., "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors," IEEE J. Sel. Areas Commun., vol. 39, no. 8, pp. 2632–2647, 2021, doi: 10.1109/JSAC.2021.3087242.

[16] M. Rajhans and V. Khawarey, "Empirical analysis of adversarial robustness and explainability drift in cybersecurity classifiers," in Proc. 18th Int. Conf. Agents and Artificial Intelligence (ICAART), vol. 4, 2026, pp. 3101–3108, doi: 10.5220/0014360100004052.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Dataset-Based Exploratory Study of Cybersecurity Threat Patterns from 2015 to 2024

Hristina Oracheva
Department of AITMIR
University of Information Science & Technology St. Paul the Apostle
Ohrid, North Macedonia
hristina.oracheva@mir.uist.edu.mk
ORCID: 0009-0005-7127-7327

Dijana Capeska Bogatinoska
Department of ISVMA
University of Information Science & Technology St. Paul the Apostle
Ohrid, North Macedonia
dijana.c.bogatinoska@uist.edu.mk
ORCID: 0000-0001-5474-8290

Daniela Mechkaroska
Department of ISVMA
University of Information Science & Technology St. Paul the Apostle
Ohrid, North Macedonia
daniela.meckaroska@uist.edu.mk
ORCID: 0000-0002-3567-8335

Abstract
Cyberattacks are becoming more frequent and technically sophisticated, posing serious risks to organizations across different sectors and regions. In this paper, we conduct an exploratory case study of cybersecurity incident information from 2015 to 2024, using exploratory data analysis, visual analytics, linear trend testing, and statistical testing. The work is built on a publicly available Kaggle dataset containing 3,000 cybersecurity incident records. The study examines temporal patterns, attack types, affected industries, derived severity levels, and temporal changes within the selected dataset. Since the dataset does not include a predefined severity variable, severity levels were derived from financial loss values and grouped into four quartiles: Low, Medium, High, and Critical. Simple linear regression was used to estimate the temporal trend in annual incident counts. At the same time, chi-square tests were applied to examine whether derived severity was associated with attack type and target industry. The results show a weak positive temporal trend, but the linear trend itself does not reach statistical significance. Within the analyzed dataset, sectors such as IT, banking, healthcare, and government appeared to have greater exposure to high-severity records. These findings highlight how useful data-driven visualizations and statistical analytics can be for cybersecurity awareness, risk interpretation, and mitigation planning. However, the conclusions should be interpreted as exploratory and dataset-dependent, rather than as a complete or comprehensive representation of real-world global cybersecurity threat patterns.

Keywords: cybersecurity threats, cyberattack trends, exploratory data analysis, visual analytics, severity analysis, cyber risk assessment, mitigation strategies

I. INTRODUCTION

The increasing dependence of modern society on digital systems has made cybersecurity a major organizational and societal concern. Organizations rely on interconnected networks to store sensitive data, perform financial transfers, and maintain essential services. Digitalization tends to increase efficiency and ease of access, but it also brings vulnerabilities that malicious users can exploit to their advantage [1]. When a cyber incident occurs, it can lead to financial losses, information leaks, service disruptions, and reputational damage [2], especially in critical areas such as finance, healthcare, government, and infrastructure [3]. Data-driven cybersecurity analytics can help uncover temporal patterns, attack categories, sector-specific exposures, and severity levels. Still, the reliability of the results largely depends on how the dataset is constructed, such as its quality and structure. Public datasets may not accurately reflect the distribution of real-world incidents, but they are still commonly used for research, visualization, and methodological demonstration. In this paper, we present an exploratory case study using a publicly available Kaggle cybersecurity dataset covering the period 2015-2024. The study conducts exploratory data analysis, visual analytics, linear trend analysis, severity classification, and chi-square testing to

examine how cybersecurity incidents emerge and evolve within the selected dataset. This study examines temporal trends in cybersecurity incident records, the exposure of different attack types and industry sectors to severe incidents, and the role of visual and statistical analytics in interpreting cybersecurity patterns. It also formulates risk-oriented mitigation recommendations based on patterns in the observed dataset. The novelty of this study lies in combining visual analytics, a derived financial-impact severity classification, linear trend analysis, and chi-square testing within a single exploratory framework for interpreting cybersecurity incidents.

The main contributions of the paper are as follows:

- Exploratory analysis of cybersecurity incident records covering the period from 2015 to 2024.
- A visual and statistical assessment of temporal trends, attack types, sectors, and derived severity levels.
- Chi-square testing, focusing on the association between derived severity level, attack type, and target industry.
- Risk-oriented mitigation recommendations, based on observed data patterns and relevant cybersecurity literature.

II. DATA ANALYSIS AND METHODOLOGY

1. Dataset Description and Preprocessing

This study uses the publicly available Kaggle dataset "Global Cybersecurity Threats 2015–2024" [4], which includes 3,000 cybersecurity incident records spanning 2015 to 2024. The dataset includes fields for country, year, attack type, target industry, financial loss in millions of dollars, number of users affected, attack source, security vulnerability type, defense mechanism, and incident resolution time. The dataset is used as the empirical data source for exploratory analysis and is not treated as independent evidence of the prevalence of real-world global cybersecurity threats. First, the structure of the dataset and variable types are examined. The records are organized by year, attack type, target industry, and country to aid in descriptive and statistical analysis. Annual incident counts and temporal patterns across attack categories were calculated by grouping the records by year and attack category. Since the data set does not include a predefined severity variable, the severity level was derived from the amount of financial loss. The values were divided into four quartile-based groups: Low, Medium, High, and Critical. These categories are intended to reflect the relative financial impact within this specific data set only. Therefore, they should not be interpreted as universal cybersecurity severity thresholds. The processed data were used for descriptive visualization, linear trend analysis, and chi-square testing. The applied statistical procedures follow standard formulations of simple linear regression, chi-square testing, expected-cell calculation, and Cramér's V effect-size estimation. The dataset is publicly available and does not come from an official incident reporting body, so the findings are considered exploratory and dependent on the dataset itself. In addition, the derived severity variable reflects only the financial impact. It does not capture the broader severity of the incident, including factors such as downtime, data sensitivity, affected users, recovery costs, and operational disruption.

2. Exploratory Data Analysis and Visualization

This subsection presents the results of the exploratory data analysis, focusing on temporal evolution, industry-level distributions, attack-type distributions, derived severity patterns, and temporal patterns across attack categories within the analyzed dataset.

3. Exploratory Linear Trend Analysis

To examine the overall time trend in the selected dataset, a simple linear regression was performed to complement the descriptive analysis of the annual cybersecurity incident records. The annual number of incident records was the dependent variable, and the year was the independent variable. The goal of this analysis was not to develop a predictive model, but rather to provide statistical support for the visual interpretation of the temporal pattern presented in Figure 1. Ordinary linear regression was used only as an exploratory summary of the annual count pattern, and the results are interpreted with caution because the analysis is based on only 10 annual observations. Between 2015 and 2024, the annual number of incident records ranged from 263 to 319. The linear regression had a positive slope of approximately 2.56 incident records per year, indicating a modest increasing trend across the dataset.

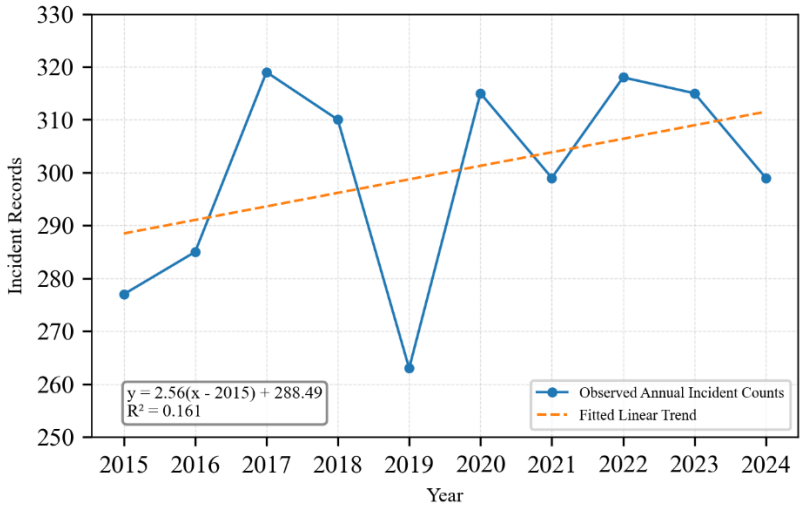


Figure 1. Annual Cybersecurity Incident Records with Exploratory Linear Trend (2015–2024)

Figure 1 presents the observed annual incident counts together with the values predicted by the fitted linear regression model. The observed counts fluctuate considerably, including a noticeable decrease in 2019 and higher values in several subsequent years. The fitted regression slope was 2.56 incident records per year, with a 95% confidence interval of [−2.21, 7.32]. The estimated trend was not statistically significant ($t(8) = 1.237$, $p = 0.251$), and the coefficient of determination was relatively low (büyük harf R kare = 0.161). Thus, this analysis does not provide sufficient statistical support for a consistent increase in annual cybersecurity incident records for the selected dataset. Although the fitted line has a positive slope, the results suggest only a rather weak positive temporal trend. This interpretation is somewhat consistent with the cybersecurity literature describing the persistence and continuous evolution of cyber threats [1], [2]. However, the results should be treated as exploratory and dataset-specific and should not be generalized to real-world global cybersecurity incident trends. Figure 2 presents the normalized percentage distribution of cybersecurity incident records across major industries and derived severity levels for the period 2015–2024. Unlike the raw-count yearly visualization, this figure summarizes the relative severity composition within each industry across the full dataset. The results indicate sector-specific variation in the proportions of Low, Medium, High, and Critical severity categories.

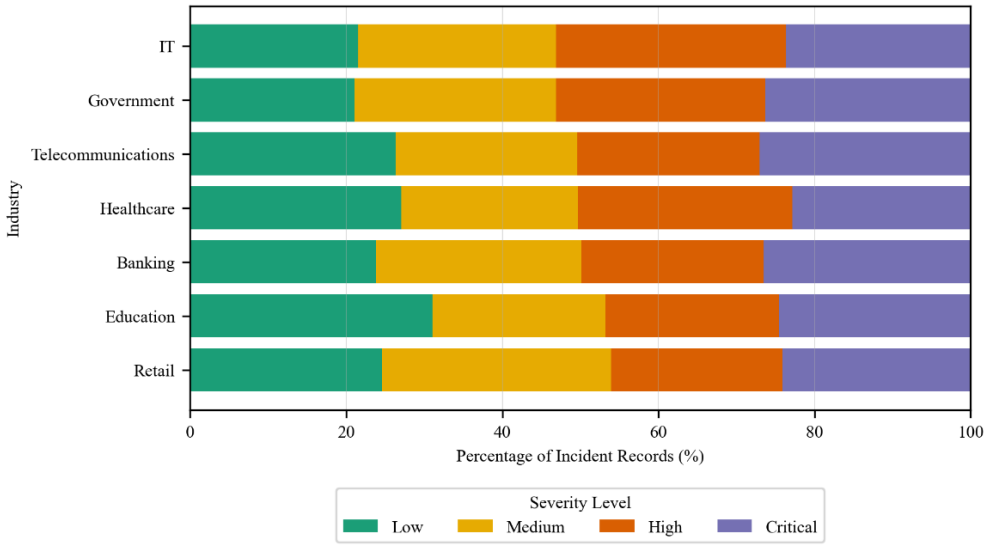


Figure 2. Normalized Distribution of Cybersecurity Incident Records by Industry and Derived Severity Level for the Period 2015–2024

These results suggest that sector-specific cybersecurity risk assessment remains important, especially for industries where cybersecurity incidents may lead to financial, operational, or public-service consequences [3], [5]. However, this interpretation is specific to the analyzed dataset because severity levels are derived from financial loss values.

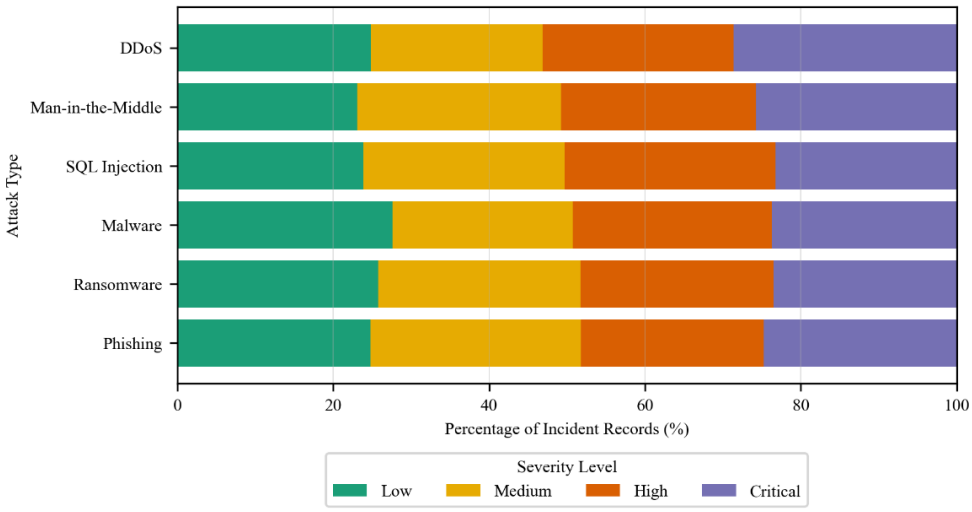


Figure 3. Normalized Distribution of Cybersecurity Incident Records by Attack Type and Derived Severity Level for the Period 2015–2024

Figure 3 presents the normalized percentage distribution of cybersecurity incident records across attack types and derived severity levels for the period 2015–2024. Unlike the raw-count yearly visualization, this figure summarizes the relative severity composition within each attack category across the full dataset. The results

indicate variation in the proportions of Low, Medium, High, and Critical severity categories across different attack types.

These results suggest that interpreting cybersecurity risk should not rely solely on the number of attack records, but also on their derived severity and potential operational impact. However, because financial loss values define severity in this study, these findings should be interpreted in the context of and with the limitations of the selected dataset.

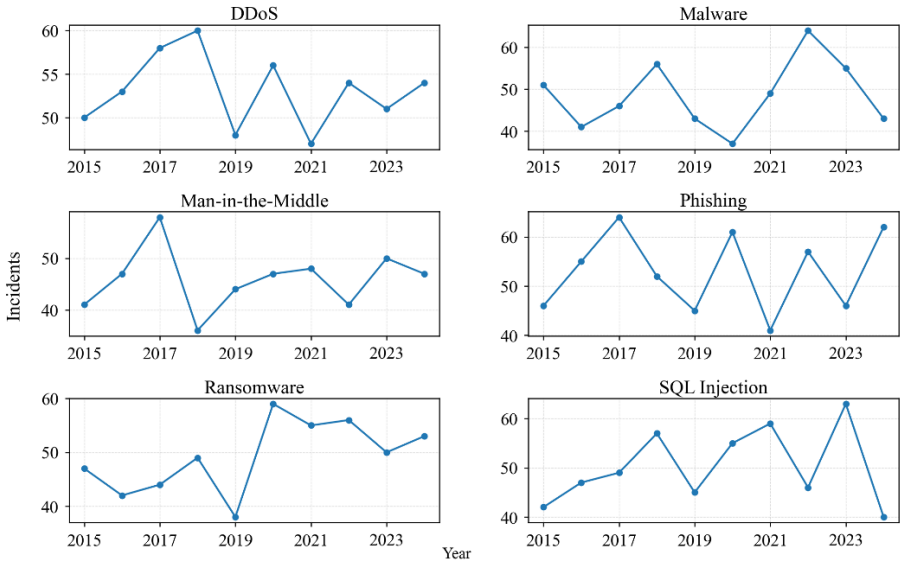


Figure 4. Temporal Evolution of Major Cyberattack Types (2015–2024)

Figure 4 presents the temporal evolution of six major cyberattack types from 2015 to 2024. The attack categories follow different yearly patterns: phishing and malware remain consistently present, while ransomware records have increased after 2020. DDoS and SQL injection show considerable fluctuations, indicating uneven temporal patterns across attack types. These findings support the importance of continuous monitoring and adaptive, multi-layered defense strategies. However, the observed patterns are dataset-specific and should be validated using real-world threat intelligence or official cybersecurity reports.

4. Statistical Analysis of Severity Patterns

To further examine severity patterns within the analyzed dataset, a chi-square test of independence was applied. The aim was to determine whether the derived severity level was statistically associated with attack type and target industry. Since the dataset did not include a predefined severity variable, severity levels were derived from financial loss values and grouped into four quartiles: Low, Medium, High, and Critical. Cross-tabulation was used to compare the distribution of derived severity levels across attack types and target industries. The chi-square test was then applied to determine whether the observed differences were statistically significant. In addition to chi-square values and p-values, degrees of freedom, sample size, Cramér's V, and expected-cell diagnostics were reported to provide a more complete statistical interpretation. The results are summarized in Table 1.

TABLE 1. CHI-SQUARE TEST RESULTS FOR SEVERITY ASSOCIATIONS

Tested relationship	χ^2	df	p-value	Cramér's V	Expected cells	Result
Attack Type vs Severity Level	12.15	15	0.667	0.037	all ≥ 5	Not significant
Target Industry vs Severity Level	30.32	18	0.034	0.058	all ≥ 5	Significant

The results show no statistically significant association between attack type and the derived severity level (χ^2 = 12.15, df = 15, p = 0.667). The corresponding Cramér's V value of 0.037 indicates a very weak association. This suggests that within the analyzed dataset, differences in severity are not strongly associated with attack categories alone. This does not mean that attack type is irrelevant in the actual assessment of cybersecurity risk. The result may be influenced by the dataset's structure, its balance, and the distribution of observations across the selected groups. There is a statistically significant relationship between the target industry and the derived severity level (χ^2 = 30.32, df = 18, p = 0.034). The corresponding Cramér's V value of 0.058 also indicates a weak association, suggesting that the statistical association should be interpreted cautiously. These results indicate that sector context may be relevant to interpreting cybersecurity risk in the analyzed dataset, though this requires validation with larger, more realistic incident datasets.

III. DISCUSSION: CYBERSECURITY INCIDENT INTELLIGENCE

The results presented in Section II suggest that, within the analyzed dataset, cybersecurity incident records remain persistent, diverse, and partly sector-dependent. This discussion interprets the key findings of the analyzed dataset in relation to relevant cybersecurity literature and selected global threat reports. The main goal is to clarify how incident frequency, derived severity levels, sector exposure, and the evolution of attack types can inform the interpretation of cybersecurity risk. However, this section remains quite dataset-specific, as the analysis relies on a publicly available dataset and derives severity levels from financial loss values.

1. Growth Trends, Severity, and Sectoral Risk

Trend analysis suggests that, within the analyzed dataset, cybersecurity incident records remained relatively high throughout the study period. Although the number of records varies from year to year, the overall pattern does not indicate a clear long-term decline. This supports the interpretation of cybersecurity threats as continuous risks rather than isolated events. However, the observed trend should be understood as dataset-specific rather than a complete representation of real-world global cybersecurity activity. Similar concerns regarding the persistence and evolution of cybersecurity threats have been discussed in global cybersecurity perspectives and threat assessment reports [6], [7]. In the dataset examined, distributed denial of service (DDoS) attacks are among the most frequent attack categories and are also associated with several high- or critical-severity levels. Sector analysis suggests that highly digitalized IT environments may be exposed to DDoS-related risks due to their high reliance on online services and digital infrastructure. Organizations that rely heavily on digital services may be more vulnerable to volumetric attacks and service disruption. Severity-based analysis shows that, within the examined dataset, IT, banking, healthcare, and government are among the sectors with higher proportions of high-severity records. Because these sectors often handle sensitive data, conduct financial operations, and support essential public services, cybersecurity incidents may have substantial consequences. The IT sector is particularly important because cybersecurity incidents affecting IT providers may have cascading effects on other organizations through supply chains and interconnected systems. Previous research on critical infrastructure also highlights the importance of adaptive and intelligent cybersecurity mechanisms [8]. From a temporal perspective, records for different attack types follow different patterns over time. Within the dataset, ransomware records increase after 2020, consistent with the broader

literature on ransomware evolution and financially motivated cybercrime [9], [10]. However, the observed trend alone is insufficient to establish a causal relationship. SQL injection and Man-in-the-Middle incident records exhibit more variable temporal patterns and may be associated with software misuse and the exploitation of network vulnerabilities. These results reinforce the need for continuous monitoring, secure software engineering, timely patch management, and adaptive detection-oriented security strategies [5], [11]. According to the findings, even in a research dataset, the interpretation of cybersecurity risk should not rely solely on incident frequency. The derived severity level, the affected sector, the geographical context, and the growth patterns of attacks allow for a more complete risk-oriented interpretation. This supports the use of data-driven cybersecurity analytics as a decision-support approach within modern risk management frameworks [5], [12]. At the same time, the results highlight the need for validation using larger, more realistic, and independently verified cybersecurity incident datasets.

IV. RISK-ORIENTED MITIGATION RECOMMENDATIONS BASED ON OBSERVED DATASET PATTERNS

Based on observed data patterns, derived severity levels, and sector-specific exposures, this section presents risk-oriented mitigation recommendations informed by the existing cybersecurity literature. These recommendations were not experimentally evaluated in this study; rather, they were derived from research findings and established cybersecurity practices. Therefore, they should be interpreted as practical recommendations rather than mitigation strategies validated by the data set. For DDoS-related risks, organizations should consider traffic filtering, cloud-based mitigation, load balancing, and network redundancy to maintain service availability [1], [13]. Phishing mitigation should combine technical filtering measures, email authentication practices, URL reputation analysis, and continuous user awareness training [2], [14]. Ransomware protection should include strict access control, offline or immutable backups, endpoint detection and response, and network segmentation as part of broader cybersecurity risk management and resilience planning [1], [5], [10]. SQL injection risks should be addressed through secure software development practices, including parameterized queries, input validation, application security testing, vulnerability scanning, and timely patch management [15]. For Man-in-the-Middle risks, organizations should enforce HTTPS and TLS, use end-to-end encryption where appropriate, and strengthen monitoring for unauthorized access points as part of broader cybersecurity risk management practices [5], [16]. Since the IT, banking, healthcare, and government sectors showed higher proportions of high-severity records in the analyzed dataset, these sectors should prioritize sector-specific risk assessment, supply chain security checks, and multi-layered cybersecurity controls [5], [6], [8].

V. CONCLUSION AND FUTURE WORK

This paper presented an exploratory analysis of a publicly available cybersecurity incident dataset spanning 2015 to 2024. The analysis used exploratory data analysis, visualization, derived severity classification, and statistical testing to examine temporal patterns, attack categories, affected sectors, and severity-related associations within the selected dataset. The exploratory linear trend analysis identified a weak positive trend, but it was not statistically significant. In the analyzed dataset, attack categories such as DDoS, phishing, malware, and ransomware appear as persistent, evolving incident records. Within the examined dataset, the sectors with higher proportions of high-severity records included IT, banking, healthcare, and government. This may be related to their dependence on digital services, sensitive data, and critical operations. However, these findings should be interpreted as dataset-specific and should not be treated as a complete representation of real-world global cybersecurity threat trends. The findings suggest that cybersecurity risk interpretation should not rely only on incident frequency. Derived severity levels, sector exposures, geographic distributions, and

temporal patterns across attack categories should also be considered. The chi-square analysis showed that, within the analyzed dataset, the target industry is significantly associated with the derived severity level, whereas attack type alone is not statistically significant. This underscores the importance of sector-specific risk assessment and calls for validation against larger, more realistic incident datasets. This study has several limitations. First, the analysis is based on a publicly available Kaggle dataset, which may not fully represent real-world cybersecurity incident distributions. Second, the severity classification was derived solely from financial loss values and does not capture other important factors, such as downtime, data sensitivity, affected users, recovery costs, or operational disruption. Third, the quartile-based severity classification is relative to the selected dataset and may lead to threshold-dependent interpretations. A full sensitivity analysis using alternative severity thresholds is left for future work due to space limitations and the exploratory scope of this conference paper. Finally, the proposed mitigation recommendations were not experimentally tested but were derived from observed patterns in the dataset and supported by existing literature. Future work should validate the findings using larger and more realistic incident datasets, official cybersecurity reports, real-time threat intelligence feeds, and sector-specific incident records [17]. Future studies should also examine the sensitivity of the severity classification by applying alternative thresholds, such as fixed financial-loss intervals, domain-informed thresholds, or composite severity scores that combine financial loss, affected users, and incident resolution time. In addition, predictive analytics and machine learning models could be explored to support more advanced cyber risk assessment.

REFERENCES

[1] ENISA, ENISA Threat Landscape 2023. Athens, Greece: European Union Agency for Cybersecurity, 2023.
[2] Verizon, 2024 Data Breach Investigations Report. New York, NY, USA: Verizon, 2024.
[3] IBM Security, Cost of a Data Breach Report 2023. Armonk, NY, USA: IBM Corporation, 2023.
[4] A. Soundankar, "Global cybersecurity threats (2015–2024)," Kaggle, dataset. [Online]. Available: https://www.kaggle.com/datasets/atharvasoundankar/global-cybersecurity-threats-2015-2024. Accessed: May 1, 2026.
[5] National Institute of Standards and Technology, Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. Gaithersburg, MD, USA: NIST, 2018, doi: 10.6028/NIST.CSWP.04162018.
[6] World Economic Forum, Global Cybersecurity Outlook 2024. Geneva, Switzerland: World Economic Forum, 2024.
[7] P. W. Singer and A. Friedman, Cybersecurity and Cyberwar: What Everyone Needs to Know. Oxford, U.K.: Oxford University Press, 2014, doi: 10.1093/wentk/9780199918096.001.0001.
[8] K. J. Raval, N. K. Jadav, T. Rathod, S. Tanwar, V. Vimal, and N. Yamsani, "A survey on safeguarding critical infrastructures: Attacks, AI security, and future directions," International Journal of Critical Infrastructure Protection, vol. 44, Art. no. 100647, 2024, doi: 10.1016/j.ijcip.2023.100647.
[9] K. Savage, P. Coogan, and H. Lau, "The evolution of ransomware," Symantec Security Response, 2015.
[10] Y. Yilmaz, O. Cetin, B. Arief, and J. Hernandez-Castro, "Investigating the impact of ransomware splash screens," Journal of Information Security and Applications, vol. 61, Art. no. 102934, 2021, doi: 10.1016/j.jisa.2021.102934.
[11] A. Bensaoud, J. Kalita, and M. Bensaoud, "A survey of malware detection using deep learning," Machine Learning with Applications, vol. 16, Art. no. 100546, Jun. 2024, doi: 10.1016/j.mlwa.2024.100546.
[12] O. Ben Fredj, A. Mihoub, M. Krichen, O. Cheikhrouhou, and A. Derhab, "Cybersecurity attack prediction: A deep learning approach," in Proc. 13th Int. Conf. Security of Information and Networks (SIN), Merkez, Turkey, Nov. 2020, pp. 1–8, doi: 10.1145/3433174.3433614.

[13] T. Kalaiselvi and M. Naveenkumar, "Comprehensive review of distributed denial-of-service attack mitigation strategies," International Journal on Science and Technology, vol. 16, no. 1, Jan.–Mar: 2025, doi: 10.71097/IJSAT.v16.i1.2917.

[14] M. Schöps, M. Gutfleisch, E. Wolter, and M. A. Sasse, "Simulated stress: A case study of the effects of a simulated phishing campaign on employees' perception, stress, and self-efficacy," in Proc. 33rd USENIX Security Symp. (USENIX Security '24), Philadelphia, PA, USA, Aug. 2024.

[15] OWASP Foundation, OWASP Top 10: The Ten Most Critical Web Application Security Risks—2025. [Online]. Available: https://owasp.org/Top10/2025/.

[16] E. Rescorla, "The Transport Layer Security (TLS) protocol version 1.3," Internet Engineering Task Force, RFC 8446, Aug. 2018. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8446. Accessed: Jun. 21, 2026.

[17] F. Cremer, B. Sheehan, M. Fortmann, A. N. Kia, M. Mullins, F. Murphy, and S. Materne, "Cyber risk and cybersecurity: A systematic review of data availability," The Geneva Papers on Risk and Insurance—Issues and Practice, vol. 47, no. 3, pp. 698–736, 2022, doi: 10.1057/s

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Ability-Based Approaches for Ambient Assisted Living in Smart Homes

Marijana Teljega
dept. of Computer Science
Kristianstad University
Kristianstad, Sweden
marijana.teljega@hkr.se
ORCID:0009-0008-7998-5055

Daniel Einarson
dept. of Computer Science
Kristianstad University
Kristianstad, Sweden
daniel.einarson@hkr.se
ORCID:0000-0002-6519-5051

Abstract
As more care is being transferred from hospitals to homes, solutions are needed to enable the care recipient's independence at home. Smart home technologies enable remote control of home devices, such as lamps and fans, via various units, including mobile phones. Depending on the degree of functional impairment the care recipient has, requirements are placed on the usability of such units. For example, reduced mobility can be compensated for with voice control. This contribution aims to study the usability of a headset for people with severe functional impairments. The contribution describes the implementation of the headsets, how they are controlled by neural activity and/or minor muscle movements, and how combinations of these, in turn, are used to control home devices.

Keywords: Brain Computer Interface, Ambient Assisted Living, Multimodal Human-Computer Interaction, Smart Home

I. INTRODUCTION

Increasingly, advanced healthcare is being delivered directly from traditional hospital settings into patients' homes [1]. In a Swedish context, this has been observed for several years [2], with reasons including a general desire to see development towards home healthcare [3]. Meanwhile, statistics show that the number of care beds in Swedish hospitals is decreasing [4]. In this context, digitalization is highlighted as a key concept for enhancing healthcare efficiency and freeing up resources [5]. Here, Telemedicine refers to IT-based systems that connect patients to medical expertise remotely [6]. In [7], it is claimed that home telemedicine shortens hospital stays, generally lowers costs, and enables faster recovery without increasing the risk of relapse. Moreover, [8] shows that Telemedicine at home increases patients' and relatives' sense of control, security, and quality of life. Furthermore, by combining telemedicine with smart home solutions – so-called Ambient Assisted Living (AAL) – patients can not only be medically monitored but also remotely control their physical home environment, such as lighting and ventilation, for safer and more independent living [9], [10]. Here [9] describes how integrating IoT (Internet of Things) with automation (controlling lighting, temperature, and fans) creates an environment in which the elderly and sick can maintain their independence despite disabilities. In this context, [11] addresses assistive IoT-based Smart Home technologies for young people with functional disabilities. That work describes a system that is under ongoing experimentation and is briefly outlined in Figure 1.

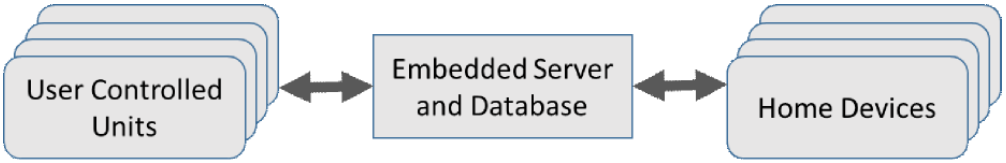


Figure 1. An overview of a Smart Home system. Users control Home Devices (lamps, fans, …) through units, such as smartphones and web interfaces. An embedded server and a database act as a mediator between units and devices [11]

To support the diverse needs of people with functional disabilities, it is especially essential to highlight user-controlled units and customize them. Thus, experiments have been performed using, for instance, cameras and accelerometers to capture gestures in the air, AI for context awareness, and speech-to-command translation. This contribution will focus on headsets (see Figure 2) as user-controlled units for communicating with Smart Home Devices. The headset captures both facial and head muscle activity (EMG) and also reads EEG activity, which is then converted into commands to control the Smart Home system.



Figure 2. OpenBCI headset, obtained from [12].

II. RELATED WORK

1. Inclusion and Sustainable Development

The principle of 'leaving no one behind' [13] is a cornerstone of the United Nations Agenda 2030. It emphasizes the need to ensure participation and quality of life for all citizens, including persons with disabilities. Several Sustainable Development Goals (SDGs) explicitly address issues related to disability, inclusion, and equal opportunities. In the context of assistive technologies and smart home environments, SDG 3 (Good Health and Well-being) and SDG 10 (Reduced Inequalities) are particularly relevant, as they promote independent living, participation in society, and equal access to supportive technologies [14]. The Convention on the Rights of Persons with Disabilities (CRPD) [15] further emphasizes the importance of enabling people with disabilities to make independent life choices and participate fully in society. To achieve these goals, research and development of accessible assistive technologies are highlighted as essential. Similarly, the World Health Organization (WHO) has stressed that access to assistive technology [16] should be regarded not only as a means of improving quality of life but also as a fundamental human right. Within this context, smart home technologies can play an important role by increasing autonomy and reducing barriers in everyday life. By providing accessible

mechanisms for environmental control, such technologies may contribute to both individual well-being and broader societal inclusion, thereby supporting the ambitions of Agenda 2030, the CRPD, and WHO initiatives on assistive technology.

2. Smart Home Technologies for People with Disabilities

AAL has become an important research area for supporting independent living through IoT-enabled smart home technologies. Recent reviews show that sensor-based monitoring, activity recognition, and health-related applications dominate the field. However, many existing solutions focus primarily on observation and monitoring rather than providing accessible mechanisms for active interaction and environmental control. Furthermore, challenges related to usability, user acceptance, and real-world deployment remain open research issues [17],[18]. Existing Ambient Assisted Living systems often rely on IoT-enabled sensing and monitoring technologies. However, recent reviews highlight persistent trade-offs between usability, privacy, cost, and system reliability. These challenges are particularly important when designing smart home solutions for individuals with severe disabilities, where accessible interaction mechanisms are essential for independent living [17],[18]. Still, while much of the AAL literature focuses on supporting aging populations, fewer studies investigate accessible smart-home interaction for users with severe physical disabilities [18] more generally. Thus, the diversity of such physical disabilities necessitates experimentation with customized, appropriate user interaction interfaces.

3. Accessible Interaction Modalities

While many smart-home solutions focus on automation and monitoring, accessible interaction technologies are needed to keep the user actively involved in controlling the environment. Different disabilities require different interaction mechanisms, motivating the need for adaptable and multimodal smart-home interfaces. Here, concepts such as Universal Design [19] and Inclusive Design [20] especially address structures and solutions to include people regardless of their abilities. Thus, accessible smart-home environments should support multiple interaction modalities rather than relying on a single interface paradigm.

Figure 1 outlines the context of the experiments on Smart Home techniques for disabled people, performed at the Computer Science department at the authors' home university. On one hand, devices of a Smart Home to interact with include lamps, fans, pull-downs, and sensors for temperature, smoke detection, and motion. On the other hand, available tools for user interaction include web-based user interfaces, smartphone apps, and speech-based controls, as shown in Figure 3 and further discussed in [11]. Thus, the architecture of Fig 3 supports the Universal Design principle of flexibility in use by providing multiple interaction mechanisms adapted to different functional abilities.

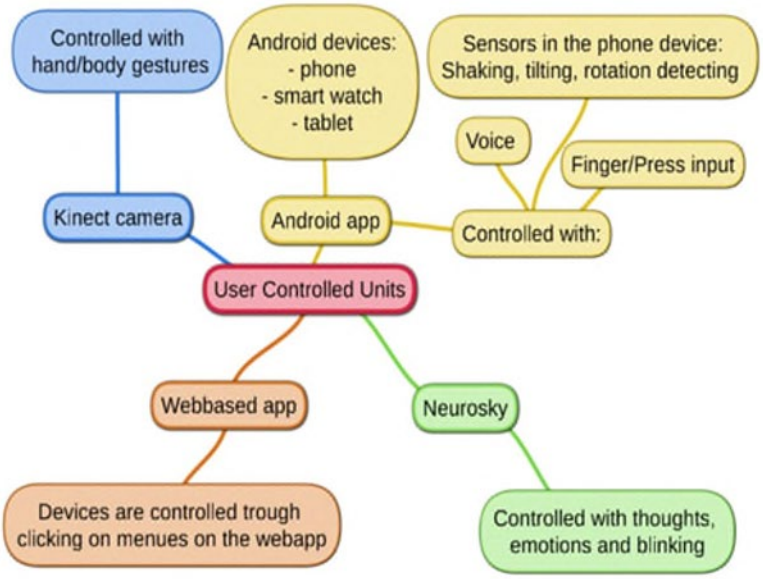


Figure 1. An overview of the Smart Home system's Users' controller units [11]

The concept of Ability-Based Design [21] is based on adapting technology to the user's abilities rather than requiring the user to adapt to the technology. Here, the term Ability-Based Design corresponds to criticism towards the previous focus on disabilities rather than abilities. This, in the context of Figure 3, would instead emphasize abilities such as: if the user can talk, then use voice-based interfaces; if the user can move their body, then use Kinect cameras [22]. If the user can use a mobile phone, use Smartphone apps; if the user can only blink, use a NeuroSky [23]. To further motivate, Figure 3 illustrates, at a conceptual level, that Multimodal Human-Computer Interaction (MHCI) emphasizes the central argument that different users need different interaction channels and that combinations of multiple modalities often provide better accessibility than a single channel [24]. Figure 3 clearly outlines several interaction options, and this contribution aims to advance interaction with Smart home devices via headsets, as discussed in the forthcoming sections. The article [25] summarizes a comparison of design possibilities for different steps in data processing, expecting that many of the current challenges will be resolved 'soon'. In our contribution, we present projects both with and without data processing. Headset manufacturers also offer APIs where signal processing and filtering have already been applied, providing users with preprocessed data. Even if modifications are possible, prepared APIs are easy to use when project time is limited, such as student projects. Instead, we focus on designing feasible command solutions suitable for AAL applications.

III. MATERIAL AND METHODS

1. Single vs Multiple Channel EEG Recording

The EEG signals obtained are weak and therefore need to be amplified to get clearer signals. This is done by placing one or more electrodes on the head or earlobes as reference points and calculating the difference between them (see Figure 2). This means that only a few electrodes are enough for measurements. But to obtain a more accurate measurement, it is desirable to use more electrodes.

Today, several headsets are available, each with a different number of electrodes for recording. In the authors' home Computer Science department, NeuroSky [26] and OpenBCI [12] with 16 electrodes were used for experimental purposes, by both students and teachers. NeuroSky works on 3,3 V (volts) and uses dry sensor technology; that is, no gel is needed to receive a signal. As presented in Figure 4, one electrode is recording activity on the forehead. This headset uses a sampling frequency of 512 Hz. One more sensor is placed on the earlobe as a reference point, and another on the ground. NeuroSky offers the possibility to read raw data and experiment with it. But it also offers a program that can translate a signal into 8 different EEG frequency bands, which can be used in further work.



Figure 4. NeuroSky MindWave headset obtained from [26]

The gel-based OpenBCI headset shown in Figure 6 enables the acquisition of 16-channel EEG data rather than 1-channel data. The Cyton board is connected to the cap and communicates via a dongle through USB or Bluetooth. Sampling frequency is 250 Hz.

2. Projects performed

The following are the four kinds of projects performed at the Computer Science department:

a) Mental exercise as a command: While stimulus can come from the outside, it can also come from the inside, for instance, through mental stimulation. Different mental exercises were designed and implemented to identify thoughts that produce the strongest effects, for future use as a command. Here [27], we used a NeuroSky headset with one electrode placed on the forehead. EEG activity in the frontal brain is associated with focused concentration, whereas relaxation (lacking concentration) is associated with lower frontal activity. Studies were conducted to identify mental exercises that put the user in the best possible state of focus. The sampling frequency was set to 512 Hz because the NeuroSky uses that frequency. Each exercise lasted 20 seconds, and each recording was repeated 5 times. 7 volunteers participated in the experiment. For more technical details, please refer to [27].

Tasks with open eyes:
“Up”- Imagine raising a window blind with your mind and watching more of the window appear.
“Down” - Imagine lowering a window blind with your mind and watching the window gradually disappear.
“Balloon” - Imagine a white balloon drifting in the wind and focus on its movement and color.
“Rectangle” - Think of one blue square. Concentrate on the blue color and follow the thick black contour line.

Tasks with closed eyes:
“Eyes Closed” - Think of whatever with eyes closed.

"Relax" - Close your eyes and relax in your muscles.
"Count Back" - Calculate backward. Start from 100.

"Point" - With closed eyes, imagine that you have a black dot between your ears in your head. Concentrate on that dot.

b) Focus, relax, and blink as a command combination: Focused and relaxed modes are again used by students [28] for command creation, combined with different numbers of blinks. Here, the NeuroSky was used with a single-electrode signal to measure its activity and control sensors on the miniature house developed by the Computer Science department (see Figure 5).

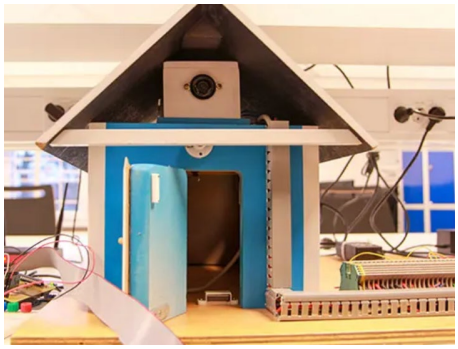


Figure 5. Miniature Smart house with varying sensors

c) Head movement combined with jaw muscle activity: Students continued to work on exploring how the creation of commands is achieved by using jaw muscle activity, EMG (electromyography- muscle response value), in combination with head movements, to construct commands. Here, the OpenBCI headset with several electrodes is used, but not all 16.

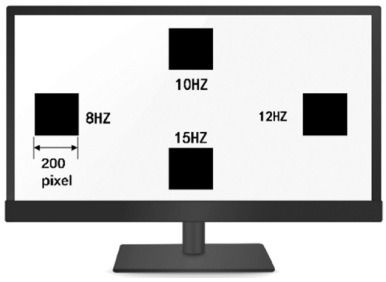


Figure 6: The image shows 4 frequencies used for the stimulus. The image is obtained from [29]

d) Using different frequencies to form a user command: Other work concerns flashing boxes at different frequencies (see Figure 6), which has been suggested for use in combination with AR (augmented reality) for future command creation. This idea is based on the content from the article [29]. Similar research has been applied to the design of BCIs for individuals who exclusively communicate with their eyes, such as when driving a wheelchair. The article [29] lies behind the motivation for this research, which showed that SSVEP (Steady State Visual Evoked Potentials) with the CCA-CWT-SVT (Canonical Correlation Analysis – Continuous

Wavelet Transformation- Super Vector Machine) algorithm achieved 91.76% accuracy in recognizing different frequencies in EEG signals.

The 4 squares with different frequencies were flashing on a screen in front of a person. Because the eyes are involved, the visual cortex is mapped to the back of the brain (occipital lobe). The electrode positions are T5, O1, O2, T6, P3, and P4 to catch the signals from this brain area. When a person looks at a flashing square with a specific frequency, for example, 15Hz, the EEG signal will show a huge spike at 15Hz. Experiments using AR (augmented reality) were added to the knowledge gained from [29] to investigate how AR can reduce the need to carry a monitor.

IV. RESULT

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or headings unless they are unavoidable.

a) The results of the experiment with mental exercises showed several things:

- Closed eyes reduce visual artifacts and give stronger results.
- Closed eyes, combined with thinking of having a point between the ears in the middle of the head, give the strongest visual effect in the spectrogram.

Finally, if combining the thought of imagining the point with counting backward, with closed eyes, the classification achieves 97% accuracy with the Nearest Neighbor classifier and the strongest visual Power Spectral Density (PSD) effect (see Figure 7).

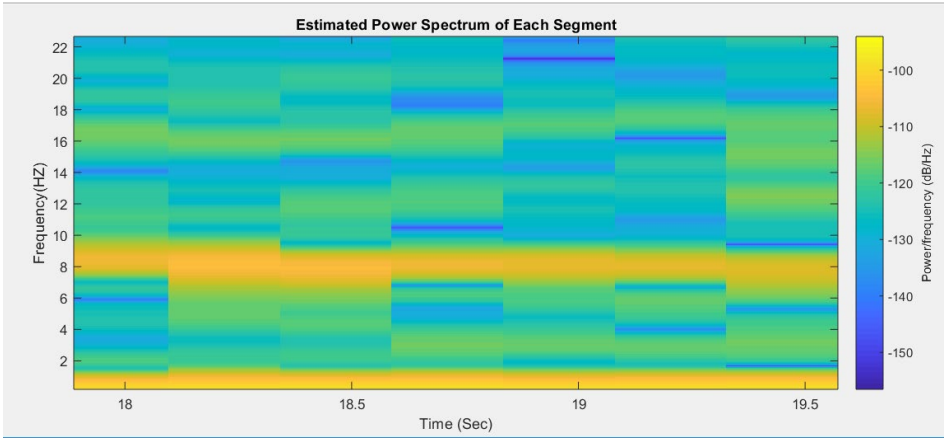


Figure 7. The mental exercise "Point" produces the strongest visual PSD effect in the spectrogram. Vertical axis: Frequency (Hz), horizontal axis: Time (Sec), bar to the right: Power/Frequency (dB/Hz)

b) The results of the experiment, focusing, relaxing, and blinking, showed that the commands are successfully used to move the mouse inside a menu represented by several boxes. You need to relax and concentrate to move

the mouse above the box, and when you are above the box, you need to blink 2 times to open it. If it fails, you will need to blink 2 times to restart (see Figure 8).

c) The result of the experiment with jaw muscles shows that muscle activity is recognized as stronger activity, which makes EMG a great combination. Also, it is important to note that the head cap is used to measure activity from the head, not from any other body part. This can be of interest to individuals, depending on their head muscle activity, because of their impairment.

d) The result of the experiment with frequencies and AR shows that the performance of AR's processors is important and that the accuracy of AR implementation depends on the implementation of AR factors. However, more work is required in this area to understand how different design factors affect accuracy.

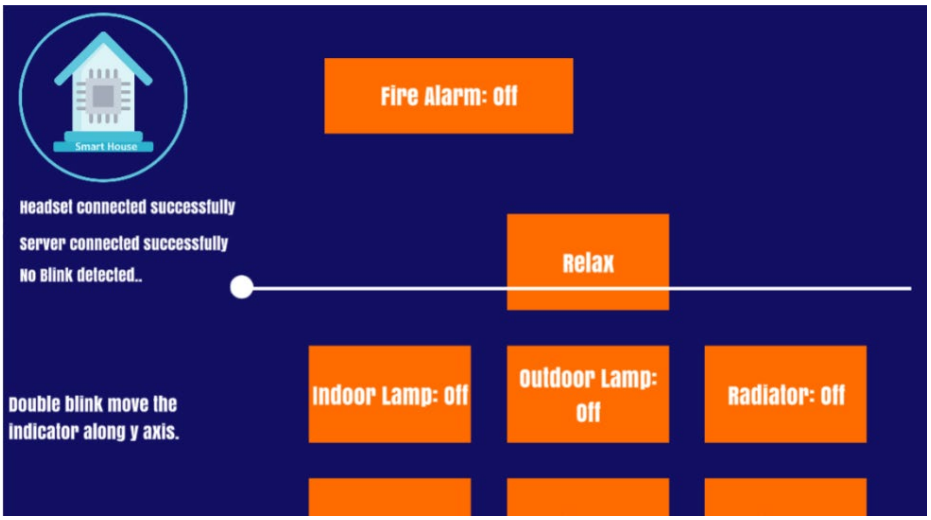


Figure 8. Smart home webpage, picture obtained from [28]

The result overview is presented in Table 1, indicating that all 4 projects show promising results and are potential candidates for AAL solution constructions.

Table 1. Result overview

Project	BCI equipment	Number of electrodes	Combinations	Raw Data Preprocessing	Applicable as an AAL solution
a)	Neurosky	one	Mental exercises	yes	yes
b)	Neurosky	one	Focus/relax + blink	no	yes
c)	OpenBCI	16	Focus+ jaw muscle	no	yes
d)	OpenBCI	16	AG + SSVEP	yes	pending

V. DISCUSSIONS

The findings indicate that both headsets have considerable potential as BCI platforms for developing command architectures that promote user independence, according to AAL and MHCI, where the user's mental health is indirectly ensured. The experimental data is not shared online due to ethical concerns. Since the studies are in an experimental phase, the headsets have not been tested by target users to confirm the results. Future work would involve continuing the work on command combination design and allowing the target user to test the prototype. The data in use is certainly of significant importance. For experimental reasons, synthetic data may be used. Moreover, involving human data raises significant ethical considerations. Furthermore, AI models and algorithms will be studied to improve the experimental results.

VI. CONCLUSIONS

The results of this study showed that brain wave activity, which was our focus from the start of the investigation, can be used to create commands that control a smart home, thereby achieving user independence. Surprisingly, small head muscles, which gave a high response in activity, are also great for designing commands. Small head muscles are also successfully combined with brain waves to increase the number of commands.

ACKNOWLEDGEMENT

The authors would like to acknowledge the valuable contributions of the students involved in projects and thesis work.

REFERENCES

[1] W. Cheng, X. Cao, W. Lian, and J. Tian, "An introduction to smart home ward–based hospital-at-home care in China," JMIR mHealth and uHealth, vol. 12, 2024, https://mhealth.jmir.org/2024/1/e44422

[2] Vårdförbundet, 'När vården flyttar hem – Den kommunala hälso- och sjukvårdens sjuksköterska i vårdens paradigmskifte' (in Swedish), 2014, accessed June 2026, nar_varden_flyttar_hem_webb.pdf

[3] Hälso- och sjukvård i hemmet – personcentrerad vård och rehabilitering, Socialstyrelsen (in Swedish), 2024, accessed June 2026, Hälso- och sjukvård i hemmet – personcentrerad vård och rehabilitering - Socialstyrelsen

[4] Sweden Hospital Beds, Trading Economics, 2022, accessed June 2026, Sweden Hospital Beds

[5] S. E. Wreile-Jensen, 'Digitalisering ska effektivisera vården och frigöra resurser' (in Swedish), Life-time – Vi följer vården, 2017, accessed June 2026, https://www.life-time.se/

[6] M. Stoltzfus, A. Kaur, A. Chawla, V. Gupta, F. N. U. Anamika, and R. Jain, "The role of telemedicine in healthcare: an overview and update," Egypt. J. Intern. Med. vol. 35, no. 49, 2023. https://doi.org/10.1186/s43162-023-00234-z

[7] N. Zychlinski et al, "Tele-medicine controlled hospital at home is associated with better outcomes than hospital stay," PLoS ONE, 2024, https://doi.org/10.1371/journal.pone.0309077

[8] K-J. Schwedler, J. Ehlers, T Ostermann, and G. Hohenberg, " The Role of Telemedicine Centers and Digital Health Applications in Home Care: Challenges and Opportunities for Family Caregivers," Healthcare, Basel, vol.14, issue.1, 2026, https://doi.org/10.3390/healthcare14010136

[9] Y. Park and J. Han, "Smart Home Advancements for Health Care and Beyond: Systematic Review of Two Decades of User-Centric Innovation," JMIR vol. 27, 2025, doi: 10.2196/62793

[10] Li KF, "Smart home technology for telemedicine and emergency management," J Ambient Intell Humaniz Comput, vol. 4, pp.535-546, 2012, https://doi.org/10.1007/s12652-012-0129-8

[11] D. Einarson and M. Teljega, “Smart Home Techniques for Young People with Functional Disabilities”, ACHI 2020, The Proceedings of the 13th International Conference on Advances in Computer-Human Interactions, pp. 294-300, ISSN: 2308-4138, ISBN: 978-1-61208-761-0, 2020, https://www.thinkmind.org/index.php?view=article&articleid=achi_2020_14_20_28015

[12] OpenBCI headset, “All-in-one EEG electrode cap bundle,” accessed June 2026, All-in-One EEG Electrode Cap Bundle – OpenBCI Shop.

[13] UNDP, “What does it mean to leave no one behind?” 2018, https://www.undp.org/publications/what-does-it-mean-leave-no-one-behind

[14] United Nations Statistics Division, “SDG Indicators,” accessed June 2026, https://unstats.un.org/sdgs/metadata/

[15] United Nations Convention on the Rights of Persons with Disabilities, “United Nations Convention on the Rights of Persons with Disabilities,” 2006, accessed June 2026, Convention on the Rights of Persons with Disabilities (CRPD) | Division for Inclusive Social Development (DISD)

[16] Global Cooperation on Assistive Technology, “Disability and rehabilitation,” World Health Organization, accessed June 2026, https://www.who.int/news-room/fact-sheets/detail/assistive-technology

[17] P. Caballero, G. Ortiz, and I. Medina-Bulo, “Systematic literature review of ambient assisted living systems supported by the Internet of Things,” Univ. Access Inf., vol. 23, pp. 1631–1656, 2024, https://doi.org/10.1007/s10209-023-01022-w

[18] B. Zieni, M. A. Ritchie, A. M. Mandalari, and F. Boem, “An Interdisciplinary Overview on Ambient Assisted Living Systems for Health Monitoring at Home: Trade-Offs and Challenges,” Sensors, vol. 25, issue 3, 2025, https://doi.org/10.3390/s25030853

[19] Center for Universal Design, “The Principles of Universal Design, “ 1997, Center for Universal Design | College of Design

[20] Microsoft Design, “Inclusive 101,” accessed June 2026, https://inclusive.microsoft.design/articles/inclusive-101-guidebook

[21] J. O. Wobbrock, S. K. Kane, K. Z. Gajos, S. Harada, and J. Froelich, “Ability-Based Design: Concept, Principles and Examples,” ACM Transactions on Accessible Computing (TACCESS), vol. 3, issue 3, no. 9, pp. 1–27, April 2021, https://doi.org/10.1145/1952383.1952384

[22] D. Levac, D. Espy, E. Fox, S. Pradhan, and J. E. Deutch, “Kinecting' With Clinicians: A Knowledge Translation Resource to Support Decision Making About Video Game Use in Rehabilitation,” Innovative Technologies Special Series, Physical Therapy, vol. 95, num. 3, pp. 426-440, 2015, https://doi.org/10.2522/ptj.20130618

[23] M. Sahu, P. Shukla, A. Chandel, S. Jain, and S. Verma, ”Eye Blinking Classification Through NeuroSky MindWave Headset Using EEGID Tool,” In: D. Gupta, A. Khanna, S. Bhattacharyya, A. E. Hassanien, S. Anand, A. Jaiswal, (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol. 1165, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-5113-0_65

[24] S. Oviatt, ”Ten myths of multimodal interaction,” Communications of the ACM, vol. 42, pp. 74-81, 1999, https://doi.org/10.1145/319382.319398

[25] S. K. Gouda, A. Choudhry, S. P. Satpathy, K. M. Shukla, A. K. Dash, and A. K. Pasayat, “Integration of EEG-based BCI technology in IoT-enabled smart home environment: An in-depth comparative analysis on human-computer interaction techniques,” Expert Systems with Applications, Elsevier, vol. 294, December 2025, https://doi.org/10.1016/j.eswa.2025.128730

[26] NeuroSky headset, “MindWave Mobile 2, ” accessed June 2026, NeuroSky MindWave Headset | EEG Brainwave Headset for Education & Training

[27] M. Teljega, “Automatic control of a window blind using EEG signals,” Master's thesis in Computer Science with Specialization in Embedded Systems, Kristianstad University, 2018, Automatic Control of a Window Blind using EEG signals - Kristianstad University Research Portal

[28] Using focused mode and blinking to turn the light on, ”Konsten att tända en lampa med tankens kraft,” accessed 2026, Konsten att tända en lampa med tankens kraft | HKR.se

[29] P. Ma et al., ”A classification algorithm of an SSVEP brain- Computer interface based on CCA fusion wavelet coefficients” Journal of Neuroscience Methods, vol. 371, April 2022, https://doi.org/10.1016/j.jneumeth.2022.109502

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Various Topological Graph Indices In Sierpinski Triangle Networks

İpek Akdoğan
Mathematics Department,
Bursa Uludag University
Bursa, Türkiye
ipekakdgnn7@gmail.com
ORCID:0009-0007-4166-2210

Hacer Özden Ayna
Mathematics Department,
Bursa Uludag University,
Bursa, Türkiye
hozden@uludag.edu.tr
ORCID:0000-0003-1556-3511

Ilker Kucukoglu
Department of Industrial Engineering,
Bursa Uludag University,
Bursa, Türkiye
ikucukoglu@uludag.edu.tr

Aysun Yurttas Gunes
Mathematics Department,
Bursa Uludag University
Bursa, Türkiye
ayurttas@uludag.edu.tr
ORCID:0000-0001-8873-1999

Aykut Sen
Department of Industrial Engineering,
Bursa Uludag University,
Bursa, Türkiye
aykutsen@uludag.edu.tr

Ismail Naci Cangul
Mathematics Department
Bursa Uludag University
Bursa, Türkiye
cangul@uludag.edu.tr
ORCID:0000-0002-0700-5774

Abstract
This paper presents a topological investigation of Sierpinski triangular networks, a classical fractal structure and a special type of graph, through their network representation, as a recursively generated figure that replicates its own structure at each iteration. The Sierpinski triangle exhibits a high level of both geometric and mathematical regularity. The study focuses on how the numbers of vertices and edges vary with the iteration level, deriving specific formulas for vertices of degrees 2 and 4. Additionally, a comprehensive set of topological graph indices, including Zagreb, Randic, Bell, Sigma, and others, is calculated and analyzed. The quantitative results obtained demonstrate that Sierpinski triangle networks are not only theoretically significant but also applicable in fields such as chemistry, communication systems, and network science. The exponential growth behavior observed with scaling aligns with the nature of complex systems, enhancing the utility of this structure in modeling large, self-similar networks. By combining fractal geometry and graph theory, this study aims to offer a fresh perspective on the analytical potential of Sierpinski-type networks.

Keywords: Sierpinski network, Sierpinski triangle, Sierpinski graph, topological graph index, modeling

I. INTRODUCTION

Fractal geometry was developed to model intricate patterns commonly found in nature, patterns that cannot be effectively explained by classical geometry. Among these, the Sierpinski triangle stands out due to its structural simplicity and strong self-similar properties. Introduced by the Polish mathematician Wacław Sierpiński, [1] this figure evolves recursively, with each iteration comprising three scaled-down replicas of the original triangle. In this paper, the Sierpinski triangle is examined using graph theory, focusing on how its iterative structure influences the distribution of vertices and edges. By categorizing vertices according to their degree, recursive and explicit-form expressions are derived for both degree 2 and degree 4 vertices. Similarly, the number and types of edges are analyzed based on the degree combinations of their connecting vertices, yielding general formulas applicable at any iteration level. Beyond these structural metrics, the study explores a variety of topological indices widely used in chemical graph theory and network analysis. Indices such as the Zagreb, Randic, Bell, and Sigma indices are computed to assess the complexity, connectivity, and irregularity of the Sierpinski triangle networks. These metrics are instrumental in evaluating molecular structures, optimizing sensor networks, and designing robust communication topologies [2]. This introductory section sets the stage

for a comprehensive analysis, wherein both theoretical derivations and applied interpretations are presented, accompanied by tabulated data illustrating the evolving behavior of the Sierpinski network across iterations.

Sierpinski triangle networks are fractal structures built according to a specific pattern. They are one of the main examples in real analysis where the notion of measure is studied. The numbers of vertices and edges in these networks vary with the construction stage k. With this notation, we will denote the Sierpinski triangle at the k-th stage by S_k. See Figure 1.

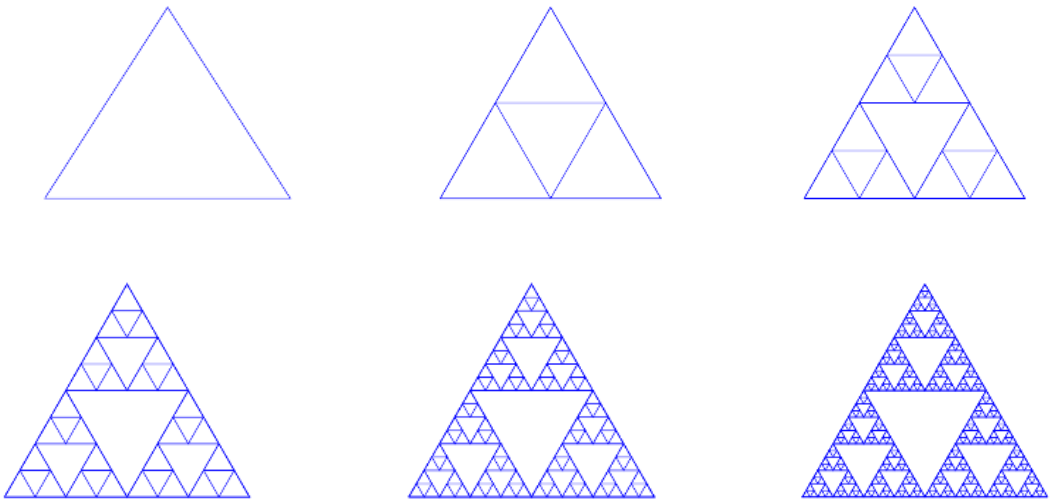


Figure 1. Sierpinski triangle networks S_k for k = 1, 2, …, 6

One of the most important reasons graphs have become one of the most popular subjects in Mathematics is that one can model almost all real-life situations with a graph and study it using mathematical tools and methods to obtain outcomes about the situation under consideration. Several important problems, such as the Traveling Salesman Problem and the Shortest Route Problem, are examples. The main tools for studying a graph model are matrices, polynomials, eigenvalues, certain parameters, and mathematical formulae called topological graph indices. When the application area is chemistry or related fields, these indices are mostly called molecular descriptors, and the study belongs to QSAR or QSPR. The definitions of various topological indices used as graph invariants, which are going to be compared in this work, are as follows:

In 1972, Gutman and Trinajstic [3] introduced the first and second Zagreb indices defined as

$$M_1(G) = \sum_{v \in V(G)} d_v^2$$

$$M_2(G) = \sum_{uv \in E(G)} dudv,$$

respectively. Here, d_i is the degree of the i-th vertex, and n is the total number of vertices. uv represents an edge connecting vertices u and v, and E is the edge set of the graph. These two indices have been studied widely for many years. For the historical background and mathematical properties of the Zagreb indices, we refer the reader to [4-8]. Another Zagreb-type index, called the forgotten index, is defined as the sum of the cubes of the degrees of all vertices in the graph, [9]:

$$F(G) = \sum_{v \in V(G)} d_v^3.$$

The hyper Zagreb index is defined as the sum of the squares of the sums of the degrees of vertices over the edges of the graph, [10], that is, as

$$HM(G) = \sum_{uv \in E(G)} (du + dv)^2.$$

It is used to characterize properties such as branching structure, molecular size, and shape. It is one of the irregularity indices that help determine the symmetrical structure of the graph under consideration. It is defined as the sum of the squares of the differences of the degrees of the vertices over all edges of the graph, [11], [12]:

$$\sigma(G) = \sum_{uv \in E(G)} (du - dv)^2.$$

The inverse sum index is defined in [13] by

$$ISI(G) = \sum_{uv \in E(G)} \frac{dudv}{du + dv}$$

and the Bell index is defined in [14] as another irregularity measure by

$$B(G) = \sum_{v \in V(G)} \left(d_v - \frac{2\mathrm{m}}{\mathrm{n}} \right)^2$$

The total irregularity index is similarly defined in [15] as yet another irregularity index by

$$Irr_t(G) = \sum_{v \in V(G)} \left| d_v - \frac{2\mathrm{m}}{\mathrm{n}} \right|.$$

The harmonic index is defined in [16] by

$$H(G) = \sum_{uv \in E(G)} \frac{2}{d_u + d_v}$$

One of the famous degree-based topological graph indices is the atom-bond connectivity index, which has molecular applications in terms of atoms and chemical bonds between them. It is defined in [17] by

$$ABC(G) = \sum_{uv \in E(G)} \sqrt{\frac{d_u + d_v - 2}{d_u d_v}}$$

The geometric-arithmetic index is defined in [18] by

$$GA(G) = \sum_{uv \in E(G)} \frac{2\sqrt{d_u d_v}}{d_u + d_v}$$

The augmented Zagreb index, which is another member of the Zagreb indices family, is defined in [19] by

$$AZ(G) = \sum_{uv \in E(G)} \left(\frac{d_u d_v}{d_u + d_v - 2}\right)^3$$

Another popular irregularity index is the Albertson index, defined in [20] by

$$Alb(G) = \sum_{uv \in E(G)} |du - dv|.$$

One of the most well-known topological graph indices is the Randic index, defined in [21] by

$$R(G) = \sum_{uv \in E(G)} \frac{1}{\sqrt{d_u d_v}}$$

The reciprocal Randic index is defined similarly to the Randic index and has some advantages in chemical calculations, see [22]:

$$RR(G) = \sum_{uv \in E(G)} \sqrt{d_u d_v}$$

1. Combinatorics of Sierpinski graphs

Up to this point, we have recalled the topological graph indices to be studied here. As seen, all the indices we are going to utilize are vertex-degree-based. Some of them are defined only in terms of each vertex degree such as the first Zagreb index, forgotten index, Bell index and total irregularity index while the others are defined

by means of the vertex degrees of end vertices of all the edges. Therefore, we next obtain the number of vertices and edges of Sk by means of vertex and edge partitions. Note that there are vertices of degree 2 and 4 only in Sk. As an immediate result, all the possible edges are of type (2,2), (2,4), or (4,4).

Observation 2.1. For k > 1, the partition of the vertices in the network Sk is as follows:

Vertex degree du	**The number of vertices of degree du**
2	3
4	$\frac{3^k-3}{2}$
Total number of vertices in S_k	$\frac{3^k+3}{2}$

The formula for the number of vertices of degree 2 is clear, as there is exactly one such vertex at each corner of the outside triangle. For the number of vertices of degree 4, the formula is obtained by the definition as a geometric series: $\sum_{i=1}^{k-1} 3^i$ and its sum is equal to $\frac{3}{2}\left(3^{k-1}-1\right)$ which can be rearranged as $\frac{3^k-3}{2}$. The total number of vertices can easily be formulized as a result of the above results.

Observation 2.2. For k > 1, the partition of the edges in the network S_k is as follows:

Vertex degrees (du,dv)	**The number of edges with end vertex degrees du and dv**
(2,2)	3 for k = 1 0 for k > 1
(2,4)	0 for k = 1 6 for k > 1
(4,4)	0 for k = 1 3^k-6 for k > 1
Total number of edges in S_k	3^k

2. Topological indices of Sierpinski triangle networks

Now, we give the results for chosen topological indices of Sierpinski triangle networks:

Theorem 3.1. The Zagreb indices of the Sierpinski triangle network S_k are

$$M_1(S_k) = 8 \cdot 3^k - 12 \quad (1)$$

$$M_2(S_k) = 16 \cdot 3^k - 48 \quad (2)$$

$$M_2(S_1) = 12 \quad (3)$$

$$F(S_k) = 32 \cdot 3^k - 72 \quad (4)$$

$$HM(S_1) = 48(5)$$

$$HM(S_k) = 64 \cdot 3^k - 168. \quad (5)$$

Proof. *Let us calculate the first Zagreb index for S_k ($k > 1$). By the vertex partition, there are three vertices of degree 2. So their contribution is 3 × 22 = 66. There are $\frac{3^k-3}{2}$ vertices of degree 4. Hence, their contribution is $\frac{3^k-3}{2} \cdot 16 = 8 \cdot 3^k - 24$. Thus, the first Zagreb index of S_k is*

$$M_1(S_k) = 12 + 8 \cdot 3^k - 24 = 8 \cdot 3^k - 12.$$

For $k = 1$, S_1 has 3 vertices of degree 2. So, $M_1(S_1) = 3 \cdot 2^2 = 12$, satisfying the rule.

Now we calculate $M_2(S_k)$. For $k = 1$, S_1 is a triangle with 3 edges, and all vertices have degree 2. Hence $M_2(S_1) = 3 \cdot (2 \cdot 2) = 12$. Let us calculate the second Zagreb index for S_k when $k > 1$. There are six edges between vertices of degree 2 and degree 4 connecting the three outer degree 2 vertices to the inner structure. Their contribution is $6 \cdot (2 \cdot 4) = 48$. Finally, the number of edges between two vertices of degree 4 is $3^k - 6$. Their contribution is $(3^k - 6) \cdot (4 \cdot 4) = 16(3^k - 6)$. Note that for $k > 1$, there are no edges directly connecting with two degree 2 vertices. Thus, for $k > 1$, the second Zagreb index is

$$M_2(S_k) = 48 + 16(3^k - 6) = 16 \cdot 3^k - 48.$$

Let us next calculate the forgotten index for S_k, $k > 1$. There are three vertices of degree 2 giving a contribution of $3 \cdot 2^3 = 24$. There are $\frac{3^k-3}{2}$ vertices of degree 4 with a contribution of $\frac{3^k-3}{2} \cdot 4^3 = 32 \cdot 3^k - 96$. Thus, the forgotten index is

$$F(S_k) = 24 + 32(3^k - 3) = 32 \cdot 3^k - 72.$$

For $k = 1$, S_1 only has three vertices of degree 2. So $F(S_1) = 3 \cdot 2^3 = 24$, satisfying the general rule.

Finally, we calculate the hyper Zagreb index of Sierpinski triangle networks starting with the particular case of S_1. As there is only one type of edge, that is, the ones of type (2,2), we have HM(S1) = 3(2 + 2) ^2 = 3(16) = 48. Secondly, for $k > 1$, we have six edges between vertices of degrees 2 and 4, giving a contribution of 6(2 + 4) ^2 = 6(36) = 216. Also there are $3^k - 6$ *edges between two vertices of degree 4 giving the second contribution of $(3^k - 6)(4 + 4)^2 = 64 \cdot 3^k - 384$. So, for $k > 1$, we find*

$$HM(S_k) = 216 + 64 \cdot 3^k - 384 = 64 \cdot 3^k - 168.$$

Theorem 3.2. The eleven degree-based topological graph indices of Sierpinski triangle networks are

$$
\begin{aligned}
&ISI(S_k) = 2\cdot 3^k - 4 && \text{for } k > 1,\\
&ISI(S_1) = 3 && \text{for } k = 1,\\
&\sigma(S_k) = 24 && \text{for } k > 1,\\
&\sigma(S_1) = 0 && \text{for } k = 1,\\
&B(S_k) = \tfrac{12(3^k-3)}{3^k+3} && \text{for } k > 1,\\
&B(S_1) = 0 && \text{for } k = 1,\\
&Irr_t(S_k) = \tfrac{12(3^k-3)}{3^k+3} && \text{for } k > 1,\\
&Irr_t(S_1) = 0 && \text{for } k = 1,\\
&H(S_k) = \tfrac{3^k+2}{4} && \text{for } k > 1,\\
&H(S_1) = \tfrac{3}{2} && \text{for } k = 1,\\
&ABC(S_k) = 3\sqrt{2}\tfrac{\sqrt{6}(3^k-6)}{4} && \text{for } k > 1,\\
&ABC(S_1) = \tfrac{3\sqrt{2}}{2} && \text{for } k = 1,\\
&GA(S_k) = 4\sqrt{2} - 6 + 3^k && \text{for } k > 1,\\
&GA(S_1) = 3 && \text{for } k = 1,\\
&AZ(S_k) = \tfrac{512\cdot 3^k-1776}{27} && \text{for } k > 1,\\
&AZ(S_1) = 24 && \text{for } k = 1,\\
&Alb(S_k) = 12 && \text{for } k > 1,\\
&Alb(S_1) = 0 && \text{for } k = 1,\\
&R(S_k) = \tfrac{6\sqrt{2}-6+3^k}{4} && \text{for } k > 1,\\
&R(S_1) = \tfrac{3}{2} && \text{for } k = 1,\\
&RR(S_k) = 12\sqrt{2} - 24 + 4\cdot 3^k && \text{for } k > 1,\\
&RR(S_1) = 6 && \text{for } k = 1.
\end{aligned}
$$

Proof. *By their definitions, all these degree based topological graph indices are calculated by means of vertex and edge partitions of S_k:*

$$ISI(S_1) = \sum_{uv\in E(S_1)} \frac{d_u d_v}{d_u + d_v} = 3\cdot\frac{2\cdot 2}{2+2} = 3\cdot\frac{4}{4} = 3$$

for k = 1. For k > 1, the contribution from 6 edges between vertices of degree 2 and 4 is $\mathbf{6}\cdot\frac{\mathbf{2\cdot 4}}{\mathbf{2+4}} = 8$ *and the contribution from* $\mathbf{3^k}-6$ *edges between vertices of degree 4 is* $(\mathbf{3^k}\text{-}6)\frac{\mathbf{4.4}}{\mathbf{4+4}} = 2(\mathbf{3^k}\text{-}6)$. *Therefore*

$$ISI(S_k) = 8 + 2(3^k - 6) = 2\cdot 3^k - 4.$$

For k=1, all three edges connect vertices of degree 2. So $\sigma(S_1) = 3 \cdot (2-2)^2 = 0$. *For k > 1, the contribution of six edges between vertices of degree 2 and 4 is* $6 \cdot (2-4)^2 = 24$ *and the contribution from* $3^k - 6$ *edges between vertices of degree 4 is* $(3^k - 6) \cdot (4-4)^2 = 0$. *So, for* $k > 1$;

$$\sigma(S_k) = 24.$$

Let us next calculate the Bell index for S_k. *For* $k = 1$, $m = 3$ *and* $n = 3$. *So* $\frac{2m}{n} = 2$. $B(S_1) = \sum_{v \in V(G)} (d_v - 2)^2 = 3(2-2)^2 = 0$. *For* $k > 1$, $m = 3^k$, $n = 3 + \frac{3}{2}(3^{k-1} - 1) = \frac{3^k+3}{2}$. *So* $\frac{2m}{n} = \frac{4 \cdot 3^k}{3^k+3}$. *There are three vertices of degree 2, and* $\frac{3}{2}(3^{k-1} - 1)$ *vertices of degree 4. Hence*

$$B(S_k) = 3\left(2 - \frac{4 \cdot 3^k}{3^k+3}\right)^2 + \frac{3}{2}(3^{k-1} - 1)\left(4 - \frac{4 \cdot 3^k}{3^k+3}\right)^2 = \frac{12(3^k-3)}{3^k+3}.$$

For the irregularity index, we can use the $\frac{2m}{n}$ *found earlier in the Bell index calculation. For* $k = 1$, $\frac{2m}{n} = 2$ *giving that*

$$Irr_t(S_1) = \sum_{u \in V(S_1)} |d_u - 2| = 3|2-2| = 0.$$

For k > 1, $\frac{2m}{n} = \frac{4 \cdot 3^k}{3^k+3}$ *and hence*

$$Irr_t(S_k) = \sum_{v \in V(S_k)} \left| d_u - \frac{4 \cdot 3^k}{3^k+3} \right| = 3\left| \frac{6 - 2 \cdot 3^k}{3^k+3} \right| + \frac{3^k-3}{2} \left| \frac{12}{3^k+3} \right|.$$

Since k > 1, $3^k > 3$ *so* $6 - 2.3^k < 0$ *and* $3^k + 3 > 0$ *giving that*

$$Irr_t(S_k) = \frac{12(3^k-3)}{3^k+3}.$$

Let us next calculate the Harmonic index for S_k. *For* $k = 1$

$$H(S_1) = \sum_{uv \in E(S_1)} \frac{2}{d_u + d_v} = 3.\frac{2}{2+2} = \frac{3}{2}.$$

For $k > 1$, the contribution of six edges between vertices of degree 2 and 4 is $6 \cdot \frac{2}{6} = 2$ and the contribution from $3^k - 6$ edges between vertices of degree 4 is $\left(3^k - 6\right) \cdot \frac{2}{8} = \frac{3k-6}{4}$ so

$$H(S_k) = 2 + \frac{3k-6}{4} = \frac{3^k + 2}{4}$$

Let us calculate atom-bond connectivity index for S_k. For $k = 1$, we have

$$ABC(S_1) = \sum_{uv \in E(S_1)} \sqrt{\frac{d_u + d_v - 2}{d_u d_v}} = 3\sqrt{\frac{2+2-2}{2.2}} = \frac{3\sqrt{2}}{2}.$$

For $k > 1$, the contribution of six edges between vertices of degree 2 and 4 is $6 \cdot \sqrt{\frac{2+4-2}{2.4}} = 3\sqrt{2}$ and the contribution from $3^k - 6$ edges between vertices of degree 4 is $\left(3^k - 6\right) \cdot \sqrt{\frac{4+4-2}{4.4}} = \left(3^k - 6\right)\frac{\sqrt{6}}{4}$. Hence

$$ABC(S_k) = 3\sqrt{2} + \frac{\sqrt{6}(3k-6)}{4}.$$

Let us calculate geometric-arithmetic index for S_k. For $k = 1$, we have

$$GA(S_1) = \sum_{uv \in E(S_1)} \frac{2\sqrt{d_u d_v}}{d_u + d_v} = 3\frac{2\sqrt{2.2}}{2+2} = 3.$$

For $k > 1$, the contribution of six edges between vertices of degree 2 and 4 is $6 \cdot \frac{2\sqrt{2.4}}{2+4} = 4\sqrt{2}$

and the contribution from $3^k - 6$ edges between vertices of degree 4 is $\left(3^k - 6\right) \cdot \frac{2\sqrt{4.4}}{4+4} = 3^k - 6$ giving

$$GA(S_k) = 4\sqrt{2} + 3^k - 6.$$

Let us calculate augmented Zagreb index for S_k. For $k = 1$, we write

$$AZ(S_1) = \sum_{uv \in E(S_1)} \left(\frac{d_u d_v}{d_u + d_v - 2}\right)^3 = 3\left(\frac{2.2}{2}\right)^3 = 24.$$

For $k > 1$ *, the contribution of six edges between vertices of degree 2 and 4 is* $6 \cdot \left(\frac{2.4}{4}\right)^3 = 48$

and the contribution of $3^k - 6$ *edges between vertices of degree 4 is* $\left(3^k - 6\right) \cdot \left(\frac{4.4}{6}\right)^3 = \frac{512}{27}\left(3^k - 6\right)$ *giving that*

$$AZ(S_k) = \sum_{uv \in E(S_1)} \left(\frac{d_u d_v}{d_u + d_v - 2}\right)^3 = 48 + \frac{512}{27}\left(3^k - 6\right) = \frac{512.3^k - 1776}{27}.$$

Let us calculate Albertson index for S_k*. For* $k = 1$*, we have*

$$Alb(S_1) = \sum_{uv \in E(S_1)} |d_u - d_v| = 3|2 - 2| = 0.$$

For $k > 1$ *, the contribution of six edges between vertices of degree 2 and 4 is* $6 \cdot |2 - 4| = 12$ *and the contribution from* $3^k - 6$ *edges between vertices of degree 4 is* $\left(3^k - 6\right) \cdot |4 - 4| = 0$ *giving that*

$$Alb(S_k) = 12.$$

Let us calculate the Randic index for S_k*. For* $k = 1$*, we get*

$$R(S_1) = \sum_{uv \in E(S_1)} \frac{1}{\sqrt{d_u d_v}} = 3.\frac{1}{\sqrt{2.2}} = \frac{3}{2}.$$

For $k > 1$ *, the contribution of six edges between vertices of degree 2 and 4 is* $6 \cdot \frac{1}{\sqrt{2.4}} = \frac{3\sqrt{2}}{2}$ *and the contribution from* $3^k - 6$ *edges between vertices of degree 4 is* $\left(3^k - 6\right) \cdot \frac{1}{\sqrt{4.4}} = \frac{3^k - 6}{4}$ *finally giving that*

$$R(S_k) = \frac{3\sqrt{2}}{2} + \frac{3^k - 6}{4} = \frac{6\sqrt{2} + 3^k - 6}{4}.$$

Let us calculate the reciprocal Randic index for S_k*. For* $k = 1$*, we have*

$$RR(S_1) = \sum_{uv \in E(S_1)} \sqrt{d_u d_v} = 3.\sqrt{2.2} = 6.$$

For $k > 1$, the contribution of six edges between vertices of degree 2 and 4 is $6 \cdot \sqrt{2.4} = 12\sqrt{2}$ and the contribution of $3^k - 6$ edges between vertices of degree 4 is $\left(3^k - 6\right) \cdot \sqrt{4.4} = 4.3^k - 24$ giving us

$$RR(S_k) = 12\sqrt{2} + 4.3^k - 24.$$

II. CONCLUSION

Networks are special graphs that interconnect people or things within a group. In this paper, 15 vertex degree based topological indices of Sierpinski triangle networks have been calculated and their general formulae are presented. These indices are important metrics for understanding the topological properties of the networks. Also, they can be used to estimate many physicochemical properties of molecular structures modeled as graphs using regression. Also, their potential applications span all fields of science, including pharmacology, industrial engineering, city planning, and neuroscience, which are not the main focus of this computational work.

REFERENCES

[1] W. Sierpinski, "Sur une courbe dont tout point est un point de ramification," Compt. Rend. Acad. Sci. Paris, vol. 160, pp. 302–305, 1915.
[2] I. Javid, H. Benish, M. Imran, A. Khan, Z. Ullah, "On some bounds of the topological indices of generalized Sierpinski and extended Sierpinski graphs," Journal of Inequalities and Applications, vol. 37, 2019.
[3] I. Gutman, N. Trinajstic, "Graph theory and molecular orbitals, Total πelectron energy of alternant hydrocarbons," Chem. Phys. Lett., vol. 17, pp. 535-538, 1972.
[4] I. Gutman, K. C. Das, "The first Zagreb index 30 years after," MATCH Commun. Math. Comput. Chem., vol. 50, pp. 83-92, 2004.
[5] S. Nikolic, G. Kovacevic, A. Milicevic, N. Trinajstic, "The Zagreb indices 30 years after," Croat. Chem. Acta, vol. 76, pp. 113-124, 2003.
[6] B. Zhou, I. Gutman, "Relations between Wiener, hyper-Wiener and Zagreb indices," Chem. Phys. Lett., vol. 394, pp. 93-95, 2004.
[7] B. Zhou, "Zagreb indices," MATCH Commun. Math. Comput. Chem., vol. 52, pp. 113-118, 2004.
[8] B. Zhou, I. Gutman, "Further properties of Zagreb indices," MATCH Commun. Math. Comput. Chem., vol. 54, pp. 233-239, 2005.
[9] B. Furtula, I. Gutman, "A forgotten topological index," J Math Chem, vol. 53, pp. 1184-1190, 2015.
[10] G.H. Shirdel, H. Rezapour, A.M. Sayadi, "The hyper-Zagreb index of graph operations," Iran. J. Math. Chem., vol. 4(2), pp. 213–220, 2013.
[11] H. Abdo, D. Dimitrov, "The total irregularity of graphs under graph operations," Miskolc Math. Notes, vol. 15, pp. 3–17, 2014.
[12] B. Furtula, I. Gutman, Z. Kovijanic Vukicevic, G. Lekishvili, G. Popivoda, "On an old/new degree–based topological index," Bull. Acad. Serbe Sci. Arts (Cl. Sci. Math. Natur.), vol. 40, pp. 19–3, 2015.
[13] D. Vukicevic, M. Gasparov, "Bond additive modeling 1. Adriatic indices," Croat. Chem. Acta, vol. 83, pp. 243-260, 2010.
[14] F. K. Bell, "A note on the irregularity of graphs," Linear Algebra Appl., vol. 161, pp. 45-54, 1992.
[15] H. Abdo, N. Cohen, D. Dimitrov, "Graphs with maximal irregularity," Filomat, vol. 28(7), pp. 1315-1322, 2014.
[16] L. Zhang, "The harmonic index for graphs," Appl. Math. Lett.,vol. 25, pp. 561-566, 2012.

[17] E. Estrada, L. Torres, L. Rodriguez, I. Gutman, "An atom-bond connectivity index: Modelling the enthalpy of formation of alkanes," Indian J. Chem., vol. 37A, pp. 849-855, 1998.
[18] D. Vukicevic, B. Furtula, "Topological index based on the ratios of geometrical and arithmetical means of end-vertex degrees of edges," J. Math. Chem., vol. 46, pp. 1369-1376, 2009.
[19] B. Furtula, A. Graovac, D. Vukicevic, "Augmented Zagreb index," J. Math. Chem., vol. 48, pp. 370-380, 2010.
[20] M. O. Albertson, "The irregularity of a graph," Ars Combin., vol. 46, pp. 219-225, 1997.
[21] M. Randic, "On characterization of molecular branching," J. Am. Chem. Soc., vol. 97, pp. 6609-6615, 1975.
[22] O. Favaron, M. Maheo, J. F. Sacle," Some eigenvalue properties of graphs (Conjectures of Grafitti-II)," Discrete Math., vol. 111, pp. 197-220, 1993.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

HeritageCare: A Predictive Maintenance And Monitoring System for Heritage Buildings

Chan Rou Mei
School of Software Engineering
TARUMT, Penang Branch
Penang, Malaysia
roumei04@gmail.com

Cho Sook Fong
Department of Computing and Information Technology
TARUMT, Penang Branch
Penang, Malaysia
chosf@tarc.edu.my

Loke Sue Voon
Department of Computing and Information Technology
TARUMT, Penang Branch
Penang, Malaysia
lokesv@tarc.edu.my

Abstract
Heritage buildings are valuable cultural assets that serve as important links to historical traditions. In Malaysia, these buildings are more than just landmarks as they define the nation's identity and play a key role in attracting tourists. However, many heritage structures face rapid deterioration due to environmental exposure, reactive maintenance practices, and a lack of modern monitoring systems. Therefore, this project proposes HeritageCare, an AI-driven system designed to support the monitoring and maintenance management of heritage buildings. The system integrates DHT22 and DS18B20 sensors to continuously collect real-time humidity and temperature data. The collected environmental data are used for real-time monitoring through the system dashboard and as input for the Long Short-Term Memory (LSTM) model to predict future maintenance dates based on environmental conditions. The model achieved 96.96% accuracy, 71.49% precision, 83.37% recall, and 75.74% F1-score. This enables a shift from reactive repairs to proactive preservation by supporting timely maintenance planning. Additionally, the platform incorporates LiDAR-based 3D modeling to ensure that restoration work preserves the original architectural design with high accuracy. The proposed system assists heritage authorities in improving maintenance planning, reducing deterioration risks, and enhancing decision-making, while also increasing public awareness and support for heritage conservation.

Keywords: Predictive Maintenance, LSTM, LiDAR 3D Modeling, Building Preservation

I. INTRODUCTION

Heritage buildings are valuable cultural assets that preserve a nation's historical identity and architectural heritage. In Malaysia, heritage sites such as George Town and Melaka contribute significantly to cultural tourism and education. However, these buildings are highly susceptible to deterioration from humidity, thermal stress, rainfall, air pollutants, wind, vibration, and human activities [1]. This study focuses solely on temperature and humidity, as they can be continuously monitored with low-cost IoT sensors to support predictive maintenance. The primary challenge in preserving these assets is the continued reliance on reactive maintenance practices, in which interventions are carried out only after visible structural damage has occurred [2]. This situation is further complicated by the limited adoption of predictive technologies in the local heritage sector, leaving authorities dependent on manual inspections that often miss early-stage risks [3], [4]. Furthermore, restoration activities often rely on two-dimensional documentation, which lacks the accuracy required to preserve complex architectural details [5], [6].

To address these challenges, this paper proposes HeritageCare, an AI-enabled heritage management platform that integrates IoT-based environmental monitoring, LSTM-based predictive maintenance, and LiDAR-based 3D modeling. Temperature and humidity data collected by DHT22 and DS18B20 sensors are monitored in real time via the system dashboard and used as inputs to predict optimal maintenance dates. Meanwhile, LiDAR-generated 3D models support accurate restoration planning by providing detailed geometric references, ensuring repair work remains consistent with the original structure. The proposed system enables data-driven heritage

conservation by facilitating early detection of deterioration, improving maintenance planning, and enhancing restoration accuracy.

II. BACKGROUND AND RELATED WORK

1. Machine Learning in Predictive Maintenance

To achieve accurate scheduling, the system uses the Long Short-Term Memory (LSTM) algorithm, a specialized type of recurrent neural network (RNN) designed for sequential data. Traditional RNNs often struggle with the "vanishing gradient" problem, which prevents them from learning long-term dependencies [7]. LSTM addresses this by using a unique cell state and a series of gates that control the flow of information over time. This makes it highly effective for learning from time-series inputs, such as sensor-captured humidity and temperature readings [8]. This capability is important for HeritageCare because the deterioration of heritage buildings is a slow process, with damage accumulating over months due to continuous exposure to weather and moisture. Unlike standard models that only look at immediate data, LSTM can remember these long-term seasonal trends and link them to current sensor readings. This allows the system to predict the "Days Remaining until Maintenance," thereby generating the predictive maintenance date and optimal timing for interventions.

2. DHT22 and DS18B20 Sensor

Humidity and temperature were selected as the primary monitoring parameters because they are the main environmental factors contributing to the deterioration of heritage buildings in tropical climates [1]. High humidity promotes moisture accumulation, leading to mold growth and wood decay, while temperature fluctuations cause thermal expansion and contraction that can produce cracks in masonry structures [9, 10]. In the proposed work, the system integrates the DHT22 and DS18B20 digital sensors with an ESP8266 microcontroller. The DHT22 is utilized for ambient monitoring, offering a 0-100% relative humidity range and ±2% accuracy [11]. Complementing this, the DS18B20 is used for temperature monitoring, measuring values from -55°C to +125°C with ±0.5°C accuracy between -10°C and +85°C [12]. Furthermore, waterproof versions of the DS18B20 provide the robustness required for monitoring temperatures within internal masonry and foundations.

3. LiDAR 3D Modeling

Unlike traditional documentation methods that rely on 2D photography or hand-drawn sketches, a 3D model captures a building's spatial depth and geometric complexity [13]. Repairs based on 2D documentation often face challenges due to reduced spatial context, because a single 2D view from limited angles cannot fully represent the building's spatial layout. In contrast, a 3D digital record provides a more accurate architectural reference for restoration. LiDAR operates by emitting laser pulses and measuring the time it takes for them to reflect off surfaces. This helps to create a dense collection of data points known as a "point cloud" [14]. This point cloud is then processed into a textured mesh, forming a realistic 3D replica. This approach is important for modeling heritage sites where complex decorative elements, detailed carvings, and irregular structural changes are difficult to document accurately using traditional 2D methods [6].

III. METHOD

1. Preliminary Study

A structured questionnaire was conducted over one week from 11 July to 18 July 2025. The survey was implemented using Google Forms. In total, 100 completed responses were received. The questionnaire was divided into four sections and contained 21 questions designed to capture user requirements comprehensively. Section A focused on gathering basic demographic information, and Section B explored "Awareness of Heritage Preservation" to assess how well people understand the importance of conserving heritage sites, the risks they face, and the benefits of protecting them. It also measured public concern about key issues such as environmental damage, physical deterioration, and insufficient maintenance. Section C examined public acceptance of modern conservation technologies. Finally, Section D focused on practical requirements such as language, device types, and design elements to ensure an accessible user experience.

2. Proposed System Architecture

Figure 1 shows the architecture diagram of the HeritageCare system. The system follows a web architecture in which the frontend and backend are decoupled and communicate via a REST API. The infrastructure is hosted across a dedicated web server and a database server to ensure scalability and security. Both the administrative tasks and public exploration features use the same centralized database for consistent data storage and retrieval.

3. Web and Server Architecture

The HeritageCare system adopts a three-tier architecture consisting of a React-based frontend, a Django backend, and a PostgreSQL database. Nginx serves the frontend, while RESTful APIs facilitate communication between the client and server. The backend manages user authentication and maintenance scheduling, whereas PostgreSQL stores heritage and maintenance. TCP-based communication ensures reliable data exchange between the application and database servers [15].

4. Sensor Integration

HeritageCare integrates DHT22 and DS18B20 sensors to monitor temperature and humidity at heritage sites in real time. Sensor readings are transmitted to an IoT gateway via the MQTT protocol, a lightweight messaging standard suitable for resource-constrained networks [16]. The gateway preprocesses and normalizes the data before forwarding it to the backend server for real-time alerts and LSTM-based predictive maintenance analysis.

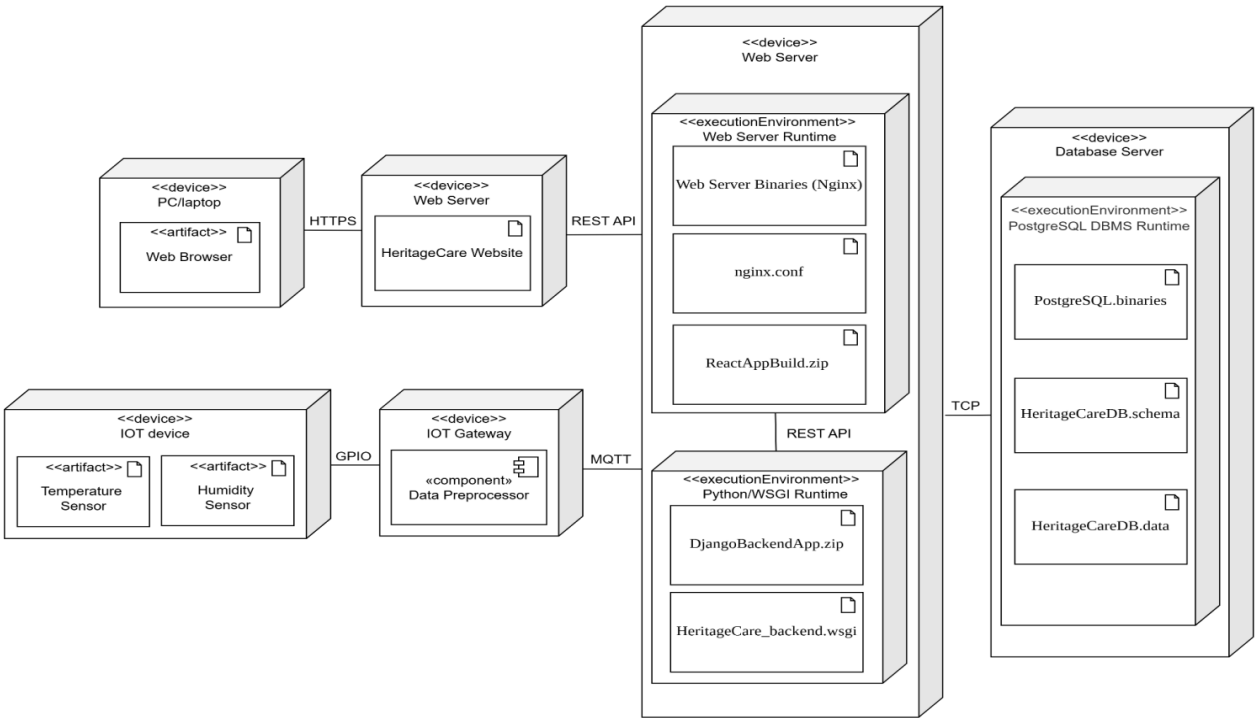


Figure 1. Architecture diagram of HeritageCare

5. System Implementation
6. Implementation of Predictive Maintenance

The predictive maintenance models were developed using an open-access environmental monitoring dataset from Zenodo [17], comprising 34,928 records with timestamps, building and weather temperature, building and weather humidity, mold risk, and thermal stress risk. The dataset was divided into 80% for training and 20% for testing. Missing values were handled using linear interpolation, while Min-Max scaling was applied to normalize the continuous features.

Three deep learning models, namely Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and One-Dimensional Convolutional Neural Network (1D-CNN), were trained to predict mold and thermal stress risks using timestamp, temperature, and humidity as input features. Model performance was evaluated using accuracy, precision, recall, and F1-score, as shown in Table 1.

Although the GRU model achieved slightly higher accuracy and F1-score, the LSTM model was selected for deployment because it provided better recall and precision. In predictive maintenance for heritage buildings, minimizing false negatives is critical, as undetected deterioration can lead to irreversible structural damage and significantly higher restoration costs. The selected LSTM model predicts mold and thermal stress risks, which are combined with the building material type to determine maintenance priority, recommend appropriate maintenance intervals, and estimate the next maintenance date.

Table 1. Performance Comparison of the Model

Final Deep Learning Comparison Matrix

Algorithm Model Architecture	Accuracy	Precision	Recall	F1-Score
LSTM (Proposed)	96.96%	**71.49%**	**83.37%**	75.74%
GRU (RNN Baseline)	**97.73%**	70.35%	76.00%	**79.41%**
1D-CNN (CNN Baseline)	94.93%	61.07%	58.29%	59.57%

7. IoT Monitoring and LiDAR Visualization

The HeritageCare system integrates IoT-based environmental monitoring and LiDAR-based 3D visualization to support proactive heritage conservation. An ESP8266 microcontroller connected to DHT22 and DS18B20 sensors collects real-time humidity and temperature data from heritage sites. It transmits the readings to the Django backend, where the collected data are processed for environmental risk assessment and maintenance date prediction using an LSTM model. Data validation and error-handling mechanisms are implemented to enhance the reliability and consistency of sensor readings. The implementation of LiDAR 3D modeling begins with capturing the heritage site using a LiDAR scanner to generate a precise point cloud representation of the structure. The captured spatial data is processed and converted into a web-compatible 3D model, with mesh optimization applied to improve model loading efficiency while preserving important architectural details. The completed model is integrated into the system via an interactive interface, allowing users to explore, zoom in on, and inspect the heritage structure from different angles.

8. System Testing

HeritageCare was evaluated using black-box and white-box testing. Black-box testing verified system functionality, including user input validation, maintenance workflow transitions, sensor monitoring, and alert generation, while white-box testing validated the Django backend and predictive maintenance logic. As shown in Tables 2 and 3, all functional requirements were satisfied, and the predictive maintenance module correctly generated maintenance schedules based on environmental conditions, building material type, and maintenance history.

Table 2. DHT22 and DS18B20 Sensors Functional Testing Results

No	Test Condition	Expected Result	Actual Result	Status
DHT22 Sensor Testing				
1	Normal room environment	Humidity is in the normal range, and the dashboard shows "Normal".	55.4% detected, dashboard shows "Normal"	Pass
2	High humidity simulation (Water sprayed near the sensor)	Humidity above 70% triggers "Abnormal" alert.	89.2% detected, "Abnormal" alert triggered	Pass
3	Boundary threshold test (30% < Humidity < 70%)	Humidity near threshold, maintain "Normal" status.	Temperature repeatedly changed between 29.9°C and 30.1°C, causing continuous status changes	Fail
4	Dry threshold test (Hairdryer used near the sensor)	Humidity below 30% triggers "Abnormal" alert.	25.3% detected, "Abnormal" alert triggered	Pass
5	Sensor disconnected from microcontroller	Detect failure, log "Sensor Error".	NaN detected, backend logged "Sensor Error"	Pass

DS18B20 Sensor Testing				
1	Normal surface temperature condition	Temperature normal, dashboard shows "Normal".	27.2°C detected, dashboard shows "Normal".	Pass
2	High temperature simulation (Hairdryer applied near the sensor)	A temperature above 40°C triggers an "Abnormal" alert.	41°C detected, "Abnormal" alert triggered	Pass
3	Boundary threshold test (20°C < Temp < 40°C)	Temperature near threshold, maintain "Normal".	39.8°C detected, dashboard shows "Normal"	Pass
4	Low temperature condition (Dipped into ice water)	A temperature below 20°C triggers an "Abnormal" alert.	19.5°C detected, "Abnormal" alert triggered	Pass
5	Sensor disconnected from microcontroller	Detect failure, log "Sensor Error".	NaN detected, backend logged "Sensor Error"	Pass

Table 3. Predictive Maintenance Testing

Environmental Condition	**Last Maintenance**	**Expected Result**	**Actual Result**	**Status**
Timber				
High Humidity + Low Thermal Stress	15 Mar 2024	Generate an earlier maintenance date due to high humidity risk	14 Apr 2024 (30 days). Reason: High humidity detected	Pass
High Thermal Stress + Low Humidity	15 Mar 2024	Generate an earlier maintenance date due to high thermal stress risk	29 Apr 2024 (45 days). Reason: High thermal stress detected	Pass
High Humidity + High Thermal Stress	15 Mar 2024	Generate the earliest maintenance date due to combined mold and thermal stress risks	29 Mar 2024 (14 days). Reason: High humidity and high thermal stress detected	Pass
Low Humidity + Low Thermal Stress	15 Mar 2024	Generate a routine maintenance date due to low environmental risk conditions.	11 Sep 2024 (180 days). Reason: Low humidity and low thermal stress detected	Pass
Masonry				
High Humidity + Low Thermal Stress	10 Jan 2024	Generate an earlier maintenance date due to high humidity risk	09 Apr 2024 (90 days). Reason: High humidity detected	Pass
High Thermal Stress + Low Humidity	10 Jan 2024	Generate an earlier maintenance date due to high thermal stress risk	09 May 2024 (120 days). Reason: High thermal stress detected	Pass
High Humidity + High Thermal Stress	10 Jan 2024	Generate the earliest maintenance date due to combined mold and thermal stress risks	10 Mar 2024 (60 days). Reason: High humidity and high thermal stress detected	Pass
Low Humidity + Low Thermal Stress	10 Jan 2024	Generate a routine maintenance date due to low environmental risk conditions.	08 Jul 2024 (180 days). Reason: Low humidity and low thermal stress detected	Pass

Brickwork				
High Humidity + Low Thermal Stress	20 Feb 2024	Generate an earlier maintenance date due to high humidity risk	20 Apr 2024 (60 days). Reason: High humidity detected	Pass
High Thermal Stress + Low Humidity	20 Feb 2024	Generate an earlier maintenance date due to high thermal stress risk	05 May 2024 (75 days). Reason: High thermal stress detected	Pass
High Humidity + High Thermal Stress	20 Feb 2024	Generate the earliest maintenance date due to combined mold and thermal stress risks	05 Apr 2024 (45 days). Reason: High humidity and high thermal stress detected	Pass
Low Humidity + Low Thermal Stress	20 Feb 2024	Generate a routine maintenance date due to low environmental risk conditions	18 Aug 2024 (180 days). Reason: Low humidity and low thermal stress detected	Pass
Metal				
High Humidity + Low Thermal Stress	05 Apr 2024	Generate an earlier maintenance date due to high humidity risk	20 May 2024 (45 days). Reason: High humidity detected	Pass
High Thermal Stress + Low Humidity	05 Apr 2024	Generate an earlier maintenance date due to high thermal stress risk	20 Apr 2024 (15 days). Reason: High thermal stress detected	Pass
High Humidity + High Thermal Stress	05 Apr 2024	Generate the earliest maintenance date due to combined corrosion and thermal stress risks	12 Apr 2024 (7 days). Reason: High humidity and high thermal stress detected	Pass
Low Humidity + Low Thermal Stress	05 Apr 2024	Generate a routine maintenance date due to low environmental risk conditions	02 Oct 2024 (180 days). Reason: Low humidity and low thermal stress detected	Pass
Concrete				
High Humidity + Low Thermal Stress	01 Dec 2023	Generate an earlier maintenance date due to high humidity risk	30 Mar 2024 (120 days). Reason: High humidity detected	Pass
High Thermal Stress + Low Humidity	01 Dec 2023	Generate an earlier maintenance date due to high thermal stress risk	30 Mar 2024 (120 days). Reason: High thermal stress detected	Pass
High Humidity + High Thermal Stress	01 Dec 2023	Generate the earliest maintenance date due to combined mold and thermal stress risks	30 Jan 2024 (60 days). Reason: High humidity and high thermal stress detected	Pass
Low Humidity + Low Thermal Stress	01 Dec 2023	Generate a routine maintenance date due to low environmental risk conditions.	29 May 2024 (180 days). Reason: Low humidity and low thermal stress detected	Pass

IV. RESULT AND DISCUSSION

1. User Acceptance Test

User Acceptance Testing (UAT) was conducted to evaluate the usability and functionality of the HeritageCare system. A total of 50 participants completed a structured questionnaire assessing the system interface, real-time environmental monitoring, predictive maintenance scheduling, and LiDAR 3D modeling using a five-point Likert scale. The evaluation results are summarised in Tables 4-6. Overall, the results indicate a high level of user acceptance. Participants rated the system positively for usability, visual consistency, and functional effectiveness. Among all features, the predictive maintenance schedule was identified as the most valuable by 50% of participants.

Table 4. UAT Response Grouping

Question	Usability	Information Quality	System Completeness	Decision Support	User Engagement
Q1. Dashboard is intuitive and easy to navigate	✓		✓		
Q2. Consistent visual style and layout	✓	✓			
Q3. Environmental visualizations are clear and easy to interpret	✓	✓	✓	✓	
Q4. Abnormal threshold notifications are effective	✓	✓	✓	✓	
Q5. Predicted maintenance schedule assists planning		✓	✓	✓	
Q6. Prediction logically aligns with environmental trends		✓	✓	✓	
Q7. 3D viewer responsiveness and control	✓		✓		✓
Q8. 3D model detail is sufficient for inspection	✓	✓	✓	✓	✓

Table 5. UAT Mean, Standard Deviation, and Percentage Scores

Indicator	Questions	Mean	Standard Deviation	Percentage
Usability	Q1, Q2, Q3, Q4, Q7, Q8	4.483	0.654	89.66%
Information Quality	Q2, Q3, Q4, Q5, Q6, Q8	4.443	0.665	88.87%
System Completeness	Q1, Q3, Q4, Q5, Q6, Q7, Q8	4.450	0.67	89.00%

Decision Support	Q3, Q4, Q5, Q6, Q8	4.444	0.67	88.88%
User Engagement	Q7, Q8	4.520	0.674	90.40%

Table 6. Interpretation of UAT Result

Indicator	Interpretation
Usability (M = 4.483, SD = 0.654)	Respondents generally agreed that the system is intuitive and easy to use. The low SD indicates good agreement among respondents.
Information Quality (M = 4.443, SD = 0.665)	The system provides clear, understandable, and useful information, with respondents showing consistent opinions.
System Completeness (M = 4.450, SD = 0.670)	Users perceived the integrated functions as comprehensive and effective in supporting heritage preservation tasks.
Decision Support (M = 4.444, SD = 0.670)	The predictive maintenance and monitoring features were rated highly, indicating that they effectively assist users in making preservation decisions.
User Engagement (M = 4.520, SD = 0.674)	This indicator achieved the highest mean score, suggesting that the interactive 3D viewer successfully encourages user participation and engagement.

2. Discussion

The User Acceptance Testing (UAT) results show that HeritageCare effectively achieves its main objectives of supporting heritage site preservation and increasing community engagement. However, user feedback and observations during system implementation identified several limitations that can be addressed in future improvements. The first limitation is that the LSTM model's performance may decrease when learning from very long sequences, as important information from earlier time steps can gradually be forgotten. As a future improvement, a Transformer-based architecture can be incorporated to better capture long-term dependencies by allowing the model to focus on the most relevant information across the entire sequence [18]. This may improve prediction accuracy and enhance the reliability of maintenance scheduling.

System testing shows that while the DHT22 and DS18B20 sensors are effective for monitoring temperature and humidity, they cannot detect other deterioration factors, such as structural shifting, rainfall, air pollutants, wind, and human activities. Future enhancements could integrate additional sensors, such as accelerometers for vibration monitoring, rain gauges for rainfall measurement, anemometers for wind monitoring, and air quality sensors for pollutant detection. These data can be incorporated as additional input features for the predictive maintenance model to provide a more comprehensive assessment of deterioration factors, thereby improving prediction accuracy and maintenance planning. In addition, the LiDAR 3D modeling process currently requires users to scan the site with a LiDAR-enabled device and manually upload the data. To improve efficiency, future versions of the system could install fixed LiDAR scanners at heritage sites to perform automated and regular 3D scans of buildings.

V. CONCLUSION

HeritageCare demonstrates a digital approach to heritage preservation by integrating environmental sensors, machine learning, and LiDAR 3D modeling into a single platform. Instead of relying on reactive maintenance, the system enables users to monitor real-time environmental conditions, view interactive 3D models of heritage structures, and utilize predictive maintenance schedules to support planning. The LSTM-based maintenance prediction model achieved 96.96% accuracy, 71.49% precision, 83.37% recall, and 75.74% F1-score, demonstrating its capability to support maintenance risk prediction. Furthermore, User Acceptance Testing involving 50 participants showed strong user acceptance, with high scores in usability (89.66%), information quality (88.87%), system completeness (89.00%), decision support (88.88%), and user engagement (90.40%). Overall, the results indicate that HeritageCare provides an effective and user-friendly platform for supporting heritage site monitoring, maintenance planning, and preservation activities, demonstrating the potential of digital technologies to enhance sustainable heritage conservation practices.

REFERENCES

[1] N. Salleh and N. H. Salleh, "A systematic review: Understanding environmental driving factors of heritage building degradation," Journal of Surveying Construction & Property, vol. 16, no. 1, pp. 95–117, Jun. 2025, doi: https://doi.org/10.22452/jscp.vol16no1.7.

[2] R. Gahrn-Andersen and M. Festila, "From Reactive to Proactive Infrastructure Maintenance: Remote Sensing Data and Practical Resilience in the Management of Leaky Pipes," Systems, vol. 12, no. 10, pp. 431–431, oct. 2024, doi: https://doi.org/10.3390/systems12100431.

[3] N. M. Zin, F. Z. Ismail, and S. Y. Said, "Sufficiency of Technology Adaptation in Risk Monitoring Criteria for Heritage Sites Management in Malaysia: Site-Based Classification," The international archives of the photogrammetry, remote sensing and spatial information sciences/International archives of the photogrammetry, remote sensing and spatial information sciences, vol. XLVIII–M9–2025, pp. 933–940, Oct. 2025, doi: https://doi.org/10.5194/isprs-archives-xlviii-m-9-2025-933-2025.

[4] N. Laohaviraphap and T. Waroonkun, "Integrating Artificial Intelligence and the Internet of Things in Cultural Heritage Preservation: A Systematic Review of Risk Management and Environmental Monitoring Strategies," Buildings, vol. 14, no. 12, pp. 3979–3979, Dec. 2024, doi: https://doi.org/10.3390/buildings14123979.

[5] G. Wu, M. Ge, Y. Wang, Y. Chen, and L. Liu, "From 2D to 3D: A Generative Model from Single Image to Digital 3D of Chinese Three Gorges Cultural Relics," Applied Sciences, vol. 16, no. 6, p. 2678, Mar. 2026, doi: https://doi.org/10.3390/app16062678.

[6] L. Zhuo, J. Zhang, and X. Hong, "Cultural heritage characteristics and damage analysis based on multidimensional data fusion and HBIM–taking the former residence of HSBC bank in Xiamen, China as an example," Heritage Science, vol. 12, no. 1, Apr. 2024, doi: https://doi.org/10.1186/s40494-024-01238-w.

[7] D. Tjøstheim, "Selected Topics in Time Series Forecasting: Statistical Models vs. Machine Learning," Entropy, vol. 27, no. 3, p. 279, Mar. 2025, doi: https://doi.org/10.3390/e27030279.

[8] S. Wang, S. Yan, D. Zhang, S. Wan, Houchen Lv, and L. Wang, "A real-time indoor temperature and relative humidity prediction method for exhibition hall based on Long Short-Term Memory network," Journal of Building Engineering, pp. 111492–111492, Dec. 2024, doi: https://doi.org/10.1016/j.jobe.2024.111492.

[9] C. Carpino, E. Loukou, F. Nicoletti, M. C. Austin, B. Andersen, and N. Arcuri, "Risk assessment of fungal formations in historic buildings based on dynamic thermo-hygrometric analysis," Acta IMEKO, vol. 12, no. 4, pp. 1–13, Dec. 2023, doi: https://doi.org/10.21014/actaimeko.v12i4.1508.

[10] A. Prosperi, P. A. Korswagen, M. Longo, and J. G. Rots, “Field observations of spatial surface temperature variations on masonry walls using infrared thermography,” Frontiers in Built Environment, vol. 12, p. 1741805, 2026, doi: https://doi.org/10.3389/fbuil.2026.1741805.

[11] Components101, "DHT22 Sensor Pinout, Specs, Equivalents, Circuit & Datasheet," Components101, Apr. 19, 2018. Available: https://components101.com/sensors/dht22-pinout-specs-datasheet. Accessed: Jul. 14, 2026.

[12] “Waterproof DS18B20 Sensor Kit,” DigiKey, 2017. Available: https://www.digikey.my/en/product-highlight/m/maxim-integrated/waterproof-ds18b20-sensor-kit?srsltid=AfmBOopI6Q6o8mDZj0Xq80GFf_sIUf_ouf2gzVH4EtAtLCmlQZC2lgR7. Accessed: Jun. 16, 2026.

[13] C. Jiang, W. Hu, and G. Zhang, “Research on fine model reconstruction method of ancient architecture based on least squares method,” npj Heritage Science, vol. 13, no. 1, Mar. 2025, doi: https://doi.org/10.1038/s40494-025-01647-5.

[14] J. Ma et al., “A review of ToF-based LiDAR,” Journal of Semiconductors, vol. 45, no. 10, p. 101201, oct. 2024, doi: https://doi.org/10.1088/1674-4926/24040015.

[15] A. Ogola, “A Survey Paper on TCP Congestion Control Algorithms,” ResearchGate, Apr. 08, 2024. Available: https://www.researchgate.net/publication/379664580_A_Survey_Paper_on_TCP_Congestion_Control_Algorithms

[16] N. S. Alotaibi, H. I. Sayed, M. Kamel, and G. F. ElKabbany, “Secure Enhancement for MQTT Protocol Using Distributed Machine Learning Framework,” Sensors, vol. 24, no. 5, pp. 1638–1638, Mar. 2024, doi: https://doi.org/10.3390/s24051638.

[17] N. Torp-Smith, “DBCDK/orholm-prediction-demo: Release with climate data,” Zenodo, May 2022, doi: https://doi.org/10.5281/zenodo.6589119.

[18] Q. Wen et al., “Transformers in Time Series: A Survey,” arXiv (Cornell University), Aug. 2023, doi: https://doi.org/10.24963/ijcai.2023/759.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

AurIA AI for University Administrative Systems

Sergio Yorick Álvarez Montalvo
Dirección de Sistemas de Información (DSI)
Universidad Autónoma de Tamaulipas (UAT)
Ciudad Victoria, Tamaulipas, México
salvarm@docentes.uat.edu.mx

Damián Ulises Cedillo Rodríguez
Dirección de Sistemas de Información (DSI)
Universidad Autónoma de Tamaulipas (UAT)
Ciudad Victoria, Tamaulipas, México
ucedillo@docentes.uat.edu.mx

Cindy Nallely Muñiz Pérez
Dirección de Sistemas de Información (DSI)
Universidad Autónoma de Tamaulipas (UAT)
Ciudad Victoria, Tamaulipas, México
cmuniz@uat.edu.mx

Abstract
This paper presents AurIA, a domain-restricted intelligent administrative support assistant developed by the Dirección de Sistemas de Información (DSI) of the Universidad Autónoma de Tamaulipas (UAT). AurIA combines a curated institutional knowledge layer with controlled access to authorized operational data through Microsoft Copilot Studio, Microsoft Power Automate, institutional REST APIs, and Microsoft SQL Server databases. The proposed three-tier architecture separates the generative artificial intelligence (AI) interface, the orchestration middleware, and the institutional data layer so that the assistant does not directly query backend databases. The knowledge layer is derived from selected current official sources and AurIA-specific information artifacts, structured according to project-defined metadata rules. A controlled internal technical evaluation comprising 205 functional and adversarial events achieved a 100% response rate, a 96.10% coherence rate, a 96.10% global effectiveness rate, and a mean response latency of 5.83 seconds. The results provide an initial technical-readiness baseline for further user-centered and longitudinal evaluation of domain-restricted generative AI assistants in university administrative environments.

Keywords: Generative AI, Microsoft Copilot Studio, conversational AI, university administrative systems, REST APIs, knowledge grounding, AI governance

I. INTRODUCTION

Digital transformation in higher education increasingly depends on institutions' ability to modernize administrative services while preserving the continuity of legacy information systems that support financial, human resources, payroll, and academic operations. Universities commonly operate heterogeneous portfolios of applications developed across multiple technological generations. Although these systems remain essential to institutional operations, users often must navigate multiple platforms, consult procedural documentation, interpret administrative requirements, and retrieve information from specialized systems. The *Universidad*

Autónoma de Tamaulipas (UAT) is a public university in northeastern Mexico with approximately 40,000 students per semester, nearly 3,000 faculty members, and administrative operations distributed across secretariats, central offices, schools, and faculties. The *Dirección de Sistemas de Información* (DSI), part of the Secretaría de Finanzas, develops, maintains, and supports institutional information systems for financial transactions, payroll processes, human resources administration, budget workflows, and related services. These systems support approximately 300 direct administrative users while indirectly affecting a broader population of up to 46,000 university stakeholders. These figures are based on internal institutional records available to the project team at the time of this study. Providing support for this environment involves more than making documentation available. Administrative information is distributed across relatively static institutional knowledge, such as organizational documents, governmental accounting regulations, system manuals, procedures, and frequently asked questions, and dynamic operational information stored in institutional systems, such as transaction status, authorization queues, balances, and other data whose availability depends on the user's role and organizational context. Based on internal DSI operational records, approximately 100 support requests are received per day through Microsoft Teams, telephone calls, and Outlook email during business hours. Many of these requests concern system navigation, procedural requirements, recurrent institutional information, and status inquiries. Traditional documentation repositories and knowledge portals can address part of the demand for support by centralizing manuals, regulations, and procedural information. However, they still require users to locate the appropriate source and interpret its contents. They cannot, by themselves, resolve requests that depend on current transactional data or user-specific authorization. Consequently, the problem addressed in this study is not simply the absence of documentation but the need for a unified, controlled mechanism for accessing both institutional knowledge and authorized operational information. To address this problem, the DSI developed AurIA, a domain-restricted intelligent administrative support assistant based on Microsoft Copilot Studio and deployed within the Microsoft cloud ecosystem. AurIA combines a curated institutional knowledge layer with controlled access to real-time information through institutional REST APIs. The knowledge layer is not limited to the direct ingestion of existing manuals; it is constructed from selected current official organizational and governmental accounting sources and complemented by AurIA-specific frequently asked questions and consolidated information artifacts structured according to the project's defined formats and metadata rules. This process improves the organization and retrieval of institutional knowledge while preserving the need for source validity and periodic content maintenance. For requests involving operational data, AurIA does not access institutional databases directly. Microsoft Power Automate acts as an orchestration layer between the conversational assistant and domain-specific REST APIs. These APIs remain responsible for applying business rules, validating user context, enforcing authorization, and returning only information within the user's permitted scope. AurIA therefore acts as a conversational access layer that combines knowledge retrieval and controlled transactional data access without transferring data-access authority to the generative AI component. The contributions of this paper are threefold. First, it presents a three-tier architecture that integrates a generative conversational interface, an orchestration middleware layer, and institutional administrative services while preventing direct AI access to backend databases. Second, it describes a knowledge-and-access model that combines curated institutional content with dynamic, permission-aware information retrieved via controlled APIs. Third, it reports a controlled internal technical evaluation comprising 205 functional and adversarial events and provides quantitative results for responsiveness, coherence, effectiveness, and response latency. The remainder of this paper is organized as follows. Section II reviews related work on conversational AI in higher education, knowledge-grounded generative systems, administrative support, and AI governance. Section III describes the AurIA architecture and institutional knowledge model. Section IV presents the administrative support capabilities included in the current scope. Section V discusses security and AI-governance considerations. Section VI describes the evaluation methodology. Section VII presents and discusses the results. Section VIII outlines limitations and future work, including user-centered and longitudinal evaluation. Section IX concludes the paper.

II. RELATED WORK

Conversational agents in higher education have been studied for information retrieval, student support, administrative assistance, and learning-related services. Pérez, Daradoumis, and Puig reported growing research interest in educational chatbots and identified diverse goals and evaluation approaches [1]. More recent reviews of generative AI chatbots in higher education describe a rapidly evolving field in which institutional adoption is accompanied by concerns about reliability, governance, and empirical validation [2]. A review focused specifically on university support chatbots shows that non-instructional use cases commonly include frequently asked questions, admissions, enrollment, and other administrative interactions [3]. Recent implementations provide closer comparisons for AurIA. Ilagan and Ilagan developed a large language model (LLM)-based virtual university support agent grounded in student handbook documents to address policy questions from students, faculty, and staff [4]. Gevher, Öncü, and Erdoğdu evaluated generative AI for large-scale learner-support inquiries and compared responses across frequently asked questions and previously unresolved cases [5]. López-Galisteo reported a 2026 university administrative case in which structured generative AI knowledge repositories and an internal assistant were evaluated against real administrative inquiries and longitudinal email-volume data [6]. These studies reinforce the relevance of knowledge-grounded conversational support and illustrate the importance of explicit evaluation criteria and institutional governance. Intelligent assistants in higher education have also been proposed as reusable architectures for accessing institution-specific information. Sajja et al. presented a platform-independent assistant architecture designed to answer domain-specific questions and reduce recurring communication barriers [7]. Although much of this literature is student-facing, it provides useful architectural and evaluation precedents for institutional assistants. Knowledge-grounded generation is particularly relevant when generated responses must remain tied to authoritative sources. Lewis et al. introduced Retrieval-Augmented Generation (RAG), which combines generation with retrieved external knowledge [8], and Siriwardhana et al. showed that retrieval-augmented systems can be adapted to specialized domains [9]. More recently, Susilo et al. reported a RAG-based university virtual assistant grounded in academic manuals and evaluated through accuracy, response time, and user satisfaction [10]. AurIA does not implement a custom RAG pipeline; instead, it uses Copilot Studio knowledge sources populated through a curated institutional preparation process. The underlying design principle is similar: generated responses to procedural questions should be grounded in a bounded, authoritative knowledge layer rather than open-domain information. Security and governance become more important when conversational systems are connected to operational administrative data. The National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF) organizes AI risk activities around the GOVERN, MAP, MEASURE, and MANAGE functions [11]. Microsoft documentation describes additional Copilot Studio controls for security, governance, data loss prevention, and knowledge source configuration [12], [13]. In this paper, Microsoft documentation is used only to describe platform-specific implementation capabilities; the research motivation and positioning are supported primarily by peer-reviewed literature. The literature, therefore, supports the use of conversational AI for university information services and the value of knowledge grounding. Still, empirical evidence remains comparatively limited for administrative assistants who combine curated institutional knowledge with permission-aware access to live transactional data through controlled enterprise APIs. AurIA addresses this practical gap by documenting an end-to-end administrative architecture, an explicit knowledge-preparation process, API-mediated authorization, and a controlled functional and adversarial technical evaluation.

III. SYSTEM ARCHITECTURE

1. Tier 1 Generative Artificial Intelligence and Interaction Layer

The top tier includes Microsoft Copilot Studio, which provides a conversational interface, topic configuration, knowledge source management, and analytics capabilities. Users interact with AurIA via natural-language prompts about UAT administrative procedures and systems. The assistant is configured to respond to questions within its institutional scope and to decline requests unrelated to its purpose. This scope control is essential because an administrative copilot should not behave as an unrestricted general-purpose assistant. AurIA uses two categories of information. Static knowledge is derived from selected current official sources, including organizational documents, governmental accounting materials, system manuals, procedural guidance, and administrative regulations. The DSI team complements these sources with AurIA-specific frequently asked questions (FAQs) and consolidated information artifacts, and applies project-defined formats and metadata rules before publishing them as knowledge sources. This curation step is intended to improve consistency and retrieval, but it does not make obsolete or inaccurate content reliable; source ownership, review, and updating remain necessary governance responsibilities. Dynamic knowledge is retrieved at query time through domain APIs when a user requests current operational data.

2. Tier 2 Integration and Orchestration Layer

The middleware tier is implemented using Microsoft Power Automate. It receives structured requests from the Copilot agent, validates the request context, and invokes the appropriate institutional service. Rather than exposing backend services directly to the conversational layer, the middleware provides a controlled point for routing and transformation. Service communication uses HTTPS and JSON payloads. Standardized request and response formats allow additional administrative domains to be integrated without changing the assistant's high-level interaction model.

3. Tier 3 Data Layer

The data layer includes UAT domain-specific REST APIs and their underlying SQL Server databases. Existing systems, such as financial, payroll, and human-resources platforms, expose controlled service endpoints that return JSON-encoded responses. The APIs are responsible for enforcing business rules, applying user-specific access restrictions, and limiting returned data to the user's authorized scope. In this model, the AI assistant does not directly query databases; it consumes only controlled service responses produced by institutional APIs. The architecture, therefore, combines curated static knowledge, dynamic data retrieval, and security-mediated service integration. Figure 1 summarizes the general design.

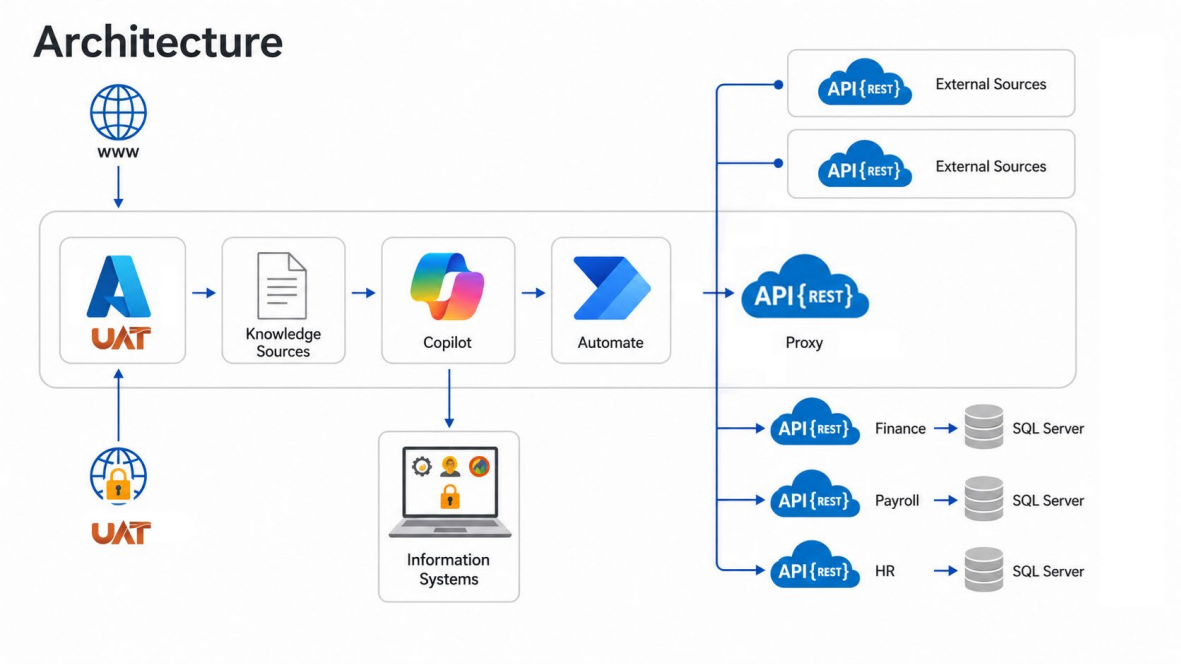


Figure 1. AurIA's three-tier architecture integrates Copilot Studio, Power Automate, REST APIs, and SQL Server data sources.

IV. ADMINISTRATIVE SUPPORT CAPABILITIES

The scope evaluated in this study centers on four administrative-support capabilities: conversational procedural assistance, authorized administrative-data retrieval, institutional information retrieval, and escalation to human support. Operational analytics are used to identify frequent topics, unresolved categories, response times, and escalation patterns. Managerial summaries and action-oriented workflows have been technically tested, and a use case is under development. Still, they are not yet deployed in the production environment and are therefore treated as future extensions rather than as evaluated core capabilities.

1. Conversational Assistance and Query Resolution

The assistant allows users to ask questions in plain Spanish about system operation, procedures, and administrative data. For procedural questions, AurIA retrieves information from the institutional knowledge layer and generates step-by-step explanations. For data-driven requests, it invokes the relevant API through Power Automate and presents the returned information in natural language. This approach reduces the need for users to know which system module, document, or menu contains the information they need. Routine inquiries include how to complete a financial form, how to verify the status of a pending authorization, where to find a payroll-related option, or which documentation applies to a specific administrative process. The present study evaluates whether AurIA can resolve representative scenarios of this type; actual reduction in DSI workload remains a hypothesis to be measured after broader deployment.

2. Authorized Administrative Data Retrieval

A representative use case involves a staff member requesting a payroll advance balance, travel credit status, reimbursement status, or the status of pending authorizations. AurIA authenticates the request context, invokes the corresponding domain API, and returns a structured response. The backend API remains responsible for enforcing user-level authorization. For example, a payroll user can access only information associated with the user's authorized institutional role, while a financial administrator can access only data permitted by the assigned organizational scope.

3. Institutional Information and Procedural Guidance

Many support requests involve static information such as institutional contacts, office locations, procedural requirements, regulations, or system navigation instructions. These requests do not require database access and can be resolved from the curated knowledge layer. The conversational interface adds contextual retrieval and synthesis to the underlying documentation, while the validity of each answer still depends on the quality and currency of the approved source material.

4. Escalation and Operational Analytics

When a query exceeds AurIA's configured resolution capacity, the assistant can initiate an escalation to human support while preserving relevant conversation context. Operational analytics can then be used to identify frequent query topics, response times, unresolved categories, and escalation patterns. Recurring unresolved questions can inform new FAQs, metadata adjustments, or updates to existing knowledge artifacts, creating a feedback loop between support activity and knowledge-base maintenance.

V. SECURITY AND Artificial Intelligence GOVERNANCE

Security is a foundational requirement for AurIA because the assistant interacts with institutional financial and administrative systems. The design combines institutional controls, API-level authentication, role-based authorization, curated knowledge governance, and cloud platform capabilities. The objective is to improve service access without expanding the information that a user is already authorized to view through institutional systems.

1. Institutional Access Controls

AurIA follows a defense-in-depth model organized around six institutional access layers. First, the user must connect from an authorized institutional network or an approved virtual private network (VPN). Second, access may be restricted by a registered device. Third, the user must authenticate with institutional credentials. Fourth, the user must have permission to use AurIA. Fifth, API responses are restricted according to the user's role and organizational unit. Sixth, confidential information requires additional authorization. These controls are intended to maintain the principle of least privilege across all interactions.

2. Service Authentication and Data Access

REST API endpoints are protected by API-key authentication and are not exposed as unrestricted public services. Requests without valid credentials are rejected before reaching backend systems. At the domain level, APIs validate the user context and return only data that corresponds to the user's permissions. Authorization is

therefore not delegated solely to the AI layer: AurIA acts as a conversational interface and orchestration client, while institutional APIs remain the authoritative enforcement points for operational data access.

3. Cloud Controls and Scope Containment

AurIA uses Microsoft Azure and Copilot Studio capabilities that support enterprise security and governance, including HTTPS communication, identity-aware access, data loss prevention, knowledge source governance, and platform-level content filtering [12], [13]. The assistant is configured as domain-restricted and is expected to decline unrelated requests, including open-domain topics, programming assistance, or sensitive subjects that fall outside its operational purpose. Scope containment is implemented through topic configuration, prompt design, bounded knowledge sources, and platform controls.

4. Alignment with the NIST Artificial Intelligence Risk Management Framework

The project does not claim formal compliance with the NIST AI RMF; however, its current design can be mapped to selected functions of the framework [11]. GOVERN is reflected in controlled knowledge-source selection, metadata rules, access responsibilities, and institutional ownership of backend services. MAP is addressed by defining the administrative domain, user roles, data sources, and out-of-scope behaviors. MEASURE is represented by the controlled functional and adversarial evaluation described in Section VI. MANAGE is supported through least-privilege API enforcement, scope containment, human escalation, knowledge maintenance, and remediation of non-conforming test cases. A formal AI-risk assessment remains future work.

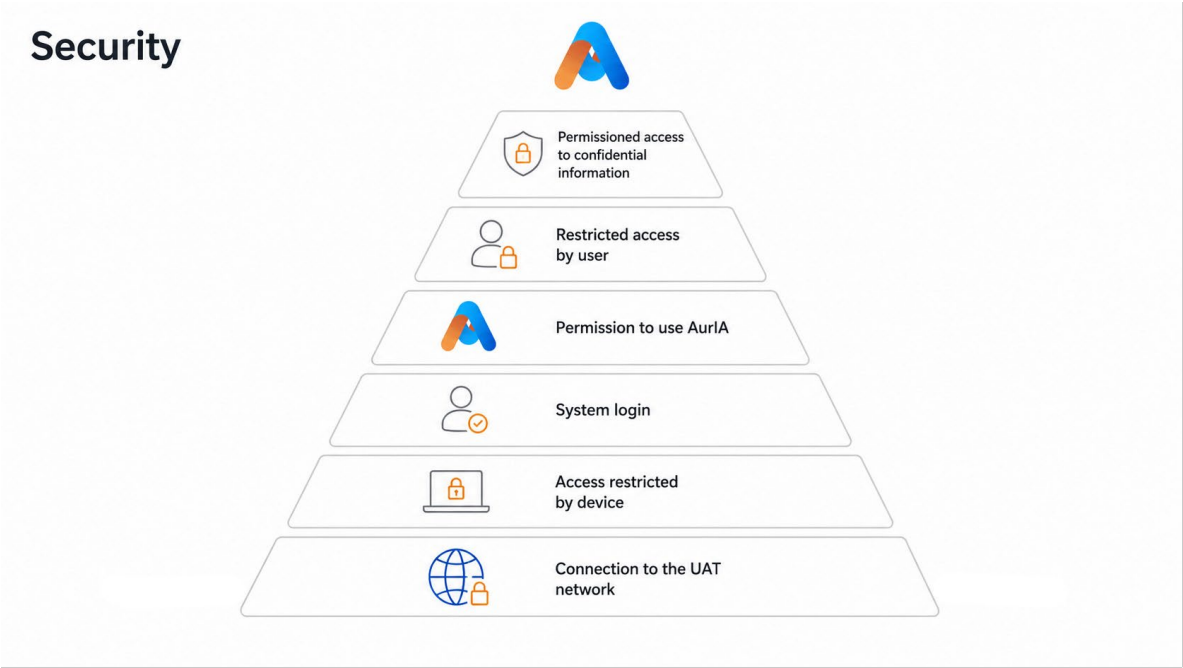


Figure 2. Institutional access-control model for AurIA, from network connectivity to permission for confidential information.

VI. EVALUATION METHODOLOGY

Before full institutional deployment, AurIA was evaluated through a controlled internal technical-readiness test designed and executed by the DSI team. The evaluation included 205 events covering representative operational queries and adversarial probes. No administrative end users participated in this evaluation. Its purpose was to determine whether the assistant could reliably produce responses, maintain coherence, fulfill user intent within its domain, and decline inappropriate requests. The test was not designed to measure usability, satisfaction, adoption, or organizational workload reduction.

1. Test Set Composition

The 205 events were constructed from representative support scenarios and risk-oriented probes. Functional events included system navigation questions, procedural guidance requests, institutional contact inquiries, and administrative data retrieval scenarios. These categories correspond to the recurring support needs described in Section I. Adversarial events included out-of-scope topic questions, code-generation requests, and sensitive-subject prompts selected to test whether the assistant maintained its intended operational scope.

2. Evaluators and Scoring Criteria

Members of the DSI development and testing team, familiar with the institutional systems and the expected behavior of AurIA, performed the evaluation. Each event was assessed using three binary criteria. Responsiveness was scored as pass when the assistant produced an output. Coherence was scored as pass when the response was grammatically understandable, internally consistent, and contextually appropriate. Effectiveness was scored as a pass when the response either correctly fulfilled the user's request within the assistant's domain or appropriately declined the request in adversarial cases. Because the evaluators were members of the development organization, the results should be interpreted as internal technical validation rather than independent user evaluation. Latency was measured as the observed round-trip time from user submission to assistant response. For requests requiring live data, this interval includes conversational processing, middleware routing, domain API invocation, SQL Server query execution, and response generation. The reported latency, therefore, represents pilot-stage performance in a controlled institutional environment rather than a guaranteed production service level.

3. Reproducibility and Interpretation

The evaluation procedure was intentionally simple so that the same events can be rerun after changes to knowledge sources, metadata, topic configuration, or API integrations. Binary scoring supports repeatable operational regression testing, but it does not capture partial correctness or user experience. Subsequent phases should add independent or end-user evaluators, inter-rater agreement, usability and satisfaction measures, task-completion metrics, and comparison with human-support baselines.

VII. RESULTS AND DISCUSSION

Table 1 summarizes the controlled technical evaluation. AurIA produced a response in all 205 events, corresponding to a response rate of 100%. Coherence and global effectiveness were both 96.10%, equivalent to 197 successful events out of 205. Eight events were classified as non-conforming and were documented for subsequent analysis and remediation through knowledge updates and topic refinement.

Table 1. AurIA controlled technical evaluation results (N = 205 events)

Total evaluation events	205
Response rate	100.00%
Coherence rate	96.10% (197 / 205)
Global effectiveness rate	96.10% (197 / 205)
Failure rate	3.90% (8 / 205)
Mean response latency	5.83 seconds
Median response latency	6.00 seconds

[a] Adversarial categories: out-of-scope topics; code-generation requests; sensitive-subject queries.

The mean response latency was 5.83 seconds, with a median of 6.00 seconds. This result demonstrates a near-immediate system response under the tested conditions. Still, it should not be treated as a direct comparison with email or telephone support, as human support channels include queueing, prioritization, availability, and issue complexity effects that were not measured in the pilot. The adversarial portion of the test indicates that AurIA maintained its configured scope during the evaluated events. Out-of-scope topics were declined or redirected, code-generation requests were not fulfilled, and sensitive-subject prompts triggered refusal behavior. Within this controlled sample, no unsafe or unrelated answer was observed for the adversarial categories. Broader red-team testing remains necessary before making stronger claims about robustness against prompt injection, social engineering, or previously unseen adversarial strategies. With respect to the research question, the pilot provides evidence that the proposed combination of curated institutional knowledge, controlled API-mediated data access, and domain restrictions can produce coherent and effective responses for representative support scenarios under internal technical testing. The 96.10% effectiveness rate is therefore an internal readiness baseline, not evidence that AurIA outperforms other university assistants or that it has already improved institutional service outcomes. The results are consistent with recent higher-education studies showing that bounded, institution-specific generative assistants can perform well when responses are grounded in curated administrative or academic knowledge [4-6], [10]. AurIA extends this pattern by adding permission-aware access to live administrative data through institutional APIs. At the same time, the eight non-conforming cases reinforce the importance of knowledge coverage, source maintenance, and regression testing after content or configuration changes.

VIII. LIMITATIONS AND FUTURE WORK

This study has several limitations. First, the evaluation sample was controlled and was executed only by DSI personnel. Although 205 events provide an initial technical-readiness baseline, they do not capture the full diversity of real-world behavior across administrative units and do not constitute a user study. Future evaluation should involve administrative personnel and measure usefulness, clarity, ease of use, satisfaction, task completion, and trust, preferably with independent assessment or inter-rater agreement where practical. A second limitation concerns organizational impact. The paper reports approximately 100 daily support requests as motivation, but the current evaluation does not demonstrate that AurIA reduces this volume, saves staff time, or improves service quality. A longitudinal deployment study should measure resolution rate, escalation frequency, support-request volume, response time, adoption, and workload before and after operational use. Recent administrative GenAI studies provide useful methodological precedents for this type of evaluation [6]. A third limitation concerns knowledge maintenance. AurIA depends on selected current official sources, AurIA-specific FAQs, consolidated information artifacts, and metadata rules. Curating these materials improves retrieval structure but does not eliminate the risk of obsolete or incorrect source information. The project, therefore, requires content ownership, review schedules, version control, update procedures, and feedback mechanisms for reporting incorrect or incomplete answers. A fourth limitation concerns compliance and security validation. This paper describes implemented and intended controls and maps them to selected NIST AI RMF functions. Still, it does not constitute a formal security audit, privacy impact assessment, or

certification of compliance. Future work should include penetration testing, prompt-injection testing, access-control verification, audit-log review, privacy analysis, and formal documentation of AI-governance responsibilities. Functional expansion will proceed only after the core support scope is validated. Managerial report summaries and controlled, action-oriented workflows have been technically tested, and one use case is under development. Still, they are not yet deployed in the production environment and were not included in the results reported here. Future studies should evaluate these capabilities separately because they involve different user roles, decision contexts, and risk levels. Additional domain APIs and knowledge sources may then be integrated using the same architecture.

IX. CONCLUSION

This paper presented AurIA, a Microsoft Copilot Studio-based assistant developed by the DSI at the Universidad Autónoma de Tamaulipas to support university administration. The proposed architecture combines a conversational AI layer, a curated institutional knowledge layer, Power Automate orchestration, and controlled REST APIs backed by SQL Server databases. The design distinguishes static institutional knowledge from dynamic operational data and keeps authorization and business-rule enforcement in institutional services rather than in the generative AI component. A controlled internal technical evaluation of 205 functional and adversarial events produced a 100% response rate, a 96.10% coherence rate, a 96.10% global effectiveness rate, and a mean latency of 5.83 seconds. These findings establish an initial technical-readiness baseline and support continued evaluation of the architecture. Still, they do not yet demonstrate user satisfaction, adoption, workload reduction, or long-term institutional impact. AurIA therefore demonstrates the technical feasibility of combining a domain-restricted generative assistant with curated institutional knowledge and permission-aware access to live administrative data through controlled APIs. The next stage of the project is user-centered and longitudinal evaluation, along with stronger security testing and formalized AI governance processes.

REFERENCES

[1] J. Q. Pérez, T. Daradoumis, and J. M. M. Puig, “Rediscovering the use of chatbots in education: A systematic literature review,” Computer Applications in Engineering Education, vol. 28, no. 6, pp. 1549–1565, 2020, doi: 10.1002/cae. 22326.

[2] C. McGrath, A. Farazouli, and T. Cerratto-Pargman, “Generative AI chatbots in higher education: A review of an emerging research area,” Higher Education, vol. 89, pp. 1533–1549, 2025, doi: 10.1007/s10734-024-01288-w.

[3] K. Peyton, S. Unnikrishnan, and B. Mulligan, “A review of university chatbots for student support: FAQs and beyond,” Discover Education, vol. 4, Art. no. 21, 2025, doi: 10.1007/s44217-025-00397-7.

[4] J. B. R. Ilagan and J. R. Ilagan, “A prototype of a conversational virtual university support agent powered by a large language model that addresses inquiries about policies in the student handbook,” Procedia Computer Science, vol. 239, pp. 1124–1131, 2024, doi: 10.1016/j.procs.2024.06.278.

[5] M. Gevher, S. E. Öncü, and E. Erdoğdu, “Exploring the potential use of generative AI for learner support in ODL at scale,” Journal of Educational Technology and Online Learning, vol. 8, no. 1, pp. 80–99, 2025, doi: 10.31681/jetol. 1559442.

[6] A. J. López-Galisteo, “Optimizing university administrative services with generative AI: Evidence from email inquiry reduction and assistant performance,” Information, vol. 17, no. 6, Art. no. 587, 2026, doi: 10.3390/info17060587.

[7] R. Sajja, Y. Sermet, D. Cwiertny, and I. Demir, "Platform-independent and curriculum-oriented intelligent assistant for higher education," International Journal of Educational Technology in Higher Education, vol. 20, Art. no. 42, 2023, doi: 10.1186/s41239-023-00412-7.

[8] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[9] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering," Transactions of the Association for Computational Linguistics, vol. 11, pp. 1–17, 2023, doi: 10.1162/tacl_a_00530.

[10] A. Susilo et al., "Enhancing student academic services through AI-driven virtual assistants using the RAG method at Universitas Terbuka," Jurnal Inovasi Teknologi Pendidikan, vol. 13, no. 1, pp. 57–69, 2026, doi: 10.21831/jitp.v13i1.83632.

[11] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2023, doi: 10.6028/NIST.AI.100-1.

[12] Microsoft, "Security and governance - Microsoft Copilot Studio," Microsoft Learn, 2026. [Online]. Available: https://learn.microsoft.com/en-us/microsoft-copilot-studio/security-and-governance. Accessed: Aug. 24, 2026.

[13] Microsoft, "Knowledge sources summary: Microsoft Copilot Studio," Microsoft Learn, 2026. [Online]. Available: https://learn.microsoft.com/en-us/microsoft-copilot-studio/knowledge-copilot-studio. Accessed: Aug. 24, 2026.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

MuseDSL: A Domain-Specific Language and Federated Framework for Explainable Museum Object Linking

Cumali Yaşar
Çanakkale Onsekiz Mart University
Çanakkale, Türkiye
cyasar@comu.edu.tr
ORCID: 0000-0002-0065-3752

Zafer Karadayı
Çanakkale Onsekiz Mart University
Çanakkale, Türkiye
karadayi@comu.edu.tr
ORCID: 0000-0002-4436-1452

Ayten Çalık
Çanakkale Onsekiz Mart University
Çanakkale, Türkiye
aytencalik@comu.edu.tr
ORCID: 0000-0002-7295-1011

Emin Ulugergerli
Çanakkale Onsekiz Mart University
Çanakkale, Türkiye
emin@comu.edu.tr
ORCID: 0000-0001-5639-1109

Abstract
Web-published museum inventories increasingly expose structured collection metadata, yet cross-institutional discovery, transparent object-to-object linking, and governed relationship management remain limited. This study presents MuseDSL, an explainable and federated framework for curator-governed museum object linking that complements existing metadata standards rather than replacing them. MuseDSL combines a federated interaction layer for decentralized discovery and metadata exchange; a domain-specific language for expressing curatorial matching rules; an agent-oriented workflow for discovery, semantic normalization, rule evaluation, and explanation generation; and a human-in-the-loop mechanism for curator validation of uncertain matches. The prototype was evaluated in a federated pilot that involved five museum inventory nodes, API-based cross-node querying, and an external mineralogical authority. The evaluation inventory contained 1,000 museum records, comprising 644 minerals, 137 rocks, 162 fossils, 39 biological specimens, and 18 other objects, from which a task pool of 100 source objects was drawn. Mineral records were submitted to Mindat for name normalization, classification checks, and evidence enrichment, resulting in a matching rate above 90% after normalization. The system also generated rule-activation traces, field-level similarity decompositions, composite scores, and decision states for candidate relationships. The results provide initial evidence that museum interoperability can be operationalized as a rule-governed, explainable, and curator-supervised interaction process without centralized metadata aggregation. MuseDSL therefore offers a practical foundation for scalable, accountable, and semantically controlled object linking across distributed museum inventories.

Keywords: museum interoperability, domain-specific language, federated systems, explainable artificial intelligence, human-in-the-loop, cultural heritage informatics

I. INTRODUCTION

Museums and cultural heritage institutions increasingly publish collection records through institutional web portals, national catalogs, and international aggregation services. This transition has improved public access to collections and created new opportunities for research, education, and digital interpretation. Nevertheless, publishing records does not automatically make distributed museum inventories interoperable with respect to object relationships. Two institutions may independently hold objects that are taxonomically related, materially similar, historically connected, or relevant to the same scientific question. At the same time, their online systems remain unable to discover or explain such relationships automatically [1], [2].

Current interoperability approaches provide strong foundations for data representation and exchange. CIDOC CRM supports semantic modeling of cultural heritage information [3], LIDO structures object metadata for publication and aggregation [4], IIIF standardizes the delivery of high-resolution digital images [5], and OAI-PMH supports repository-level metadata harvesting [6], [7]. Web-oriented vocabularies and structured data approaches further improve machine readability and discoverability [8]. These contributions are essential, but they primarily address description, publication, harvesting, or aggregation. They do not fully specify how autonomous museum systems should discover related objects, evaluate candidate links, explain why a link is proposed, and allow experts to govern that link over time. This gap is particularly important for natural history and scientific collections. When a fossil, mineral, rock, biological specimen, or archaeological object is registered in one museum, related or comparable objects may already be held by other institutions. A curator may therefore ask a practical question: which museums hold objects that belong to the same taxonomic group, geological period, material class, provenance context, or scientifically relevant neighborhood as this new record? A useful computational system should not only return candidate objects; it should also expose the evidence behind each suggestion and preserve the authority of museum professionals to validate or reject the relationship. This paper presents MuseDSL, a federated and explainable framework for cross-institutional linking of museum objects. MuseDSL combines a lightweight Museum Interoperability Protocol, a domain-specific language for curatorial matching rules, task-oriented software agents, a shared operational representation named M-CORE, and a human-in-the-loop governance process. The framework does not require participating museums to move their complete inventories into a central repository. Instead, each institution remains an autonomous node that can expose selected metadata, receive federated queries, return candidate objects, and participate in the controlled formation of relationships. This conference version focuses on the architecture, rule model, operational workflow, and initial federated pilot. The extended journal version is intended to include broader baseline comparisons, ablation experiments, threshold sensitivity studies, and larger-scale curator-centered evaluation. The present contribution is therefore positioned as a concise but complete description of the core framework and its initial empirical feasibility.

II. RELATED WORK AND RESEARCH GAP

Semantic interoperability in cultural heritage has traditionally been approached through conceptual models, metadata schemas, and repository protocols. CIDOC CRM provides a shared ontological framework for representing events, actors, places, objects, and provenance relations [3]. LIDO provides a practical XML-based schema for exchanging descriptive museum records [4]. IIIF supports interoperable access to visual resources [5], whereas OAI-PMH enables low-barrier metadata harvesting across distributed repositories [6], [7]. These mechanisms address important parts of the interoperability problem, but they do not define a complete operational process for accountable object-to-object linking between autonomous institutions. Linked data and structured web publication improve the exposure of cultural heritage resources to external systems and search engines [8]. However, publishing semantically enriched records is different from governing a proposed relationship. A relationship such as the same specimen type, a close taxonomic match, a related geological context, or a provenance association carries interpretive meaning. It may need to be reviewed, revised, accepted, or withdrawn. This requirement introduces a governance layer that is not normally central to harvesting or aggregation protocols. Agent-based architectures are suitable for separating discovery, transformation, matching, and coordination tasks in distributed environments [1]. In MuseDSL, the term agent is used in a software architectural sense: specialized modules perform defined tasks rather than claiming emergent autonomous intelligence. Explainable AI research also motivates the requirement that automated decisions should be inspectable, especially when they support expert judgment [9]. Domain-specific languages offer a complementary mechanism by expressing expert rules in a form that can be inspected and executed by software [10]. Within the cultural heritage domain itself, agent-oriented and ontology-driven systems have so

far been applied mainly to visitor interaction and content exploration, for example, ontology-based chatbots for experiential learning [11] and conversational agents for navigating European aggregated collections [12], rather than to the governance of links between institutional records. Earlier work on semantic bridges between museums, libraries, and archives likewise emphasized shared conceptual models rather than an operational linking process [13]. MuseDSL integrates these ideas into a single framework: federated discovery, domain rules, explainable evidence, and curator-governed relationship lifecycle management.

Table 1. Positioning of MuseDSL relative to common interoperability approaches

Approach	Main purpose	Federated discovery	Object linking	Explanation	Human review
CIDOC CRM	Semantic modelling	No	Conceptual	Limited	Conceptual
LIDO	Metadata exchange	No	No	No	No
IIIF	Image interoperability	No	No	No	No
OAI-PMH	Metadata harvesting	Limited	No	No	No
MuseDSL	Governed object linking	Yes	Yes	Yes	Yes

III. MUSEDSL FRAMEWORK ARCHITECTURE

The proposed framework models each participating museum as an autonomous digital node.

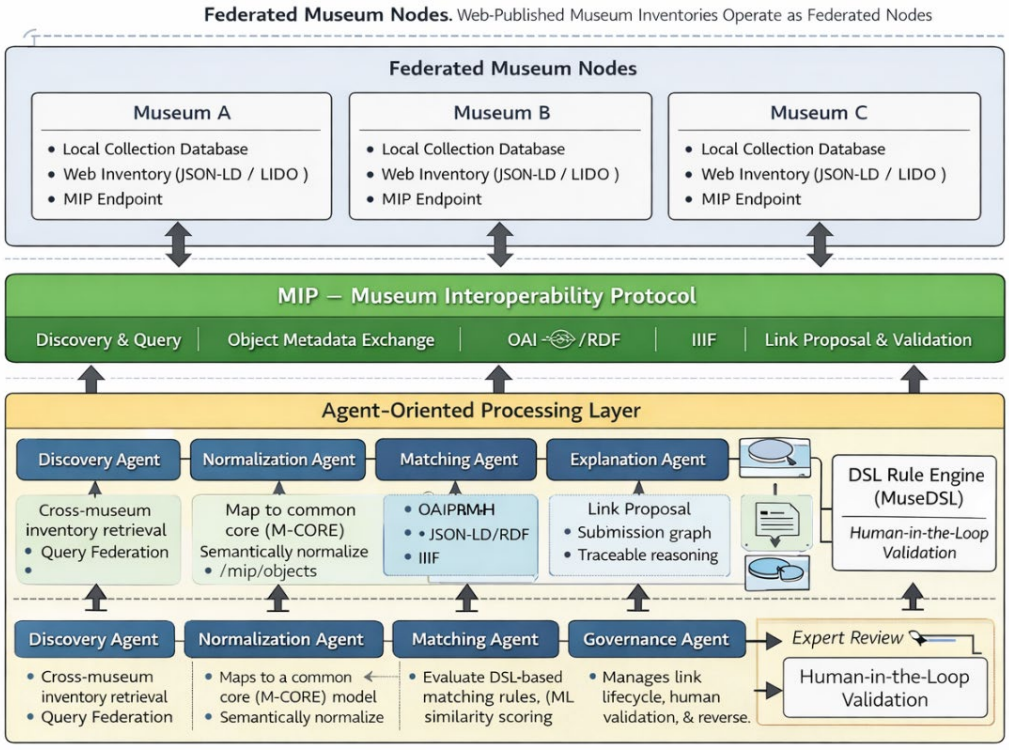


Figure 1. Layered architecture of the MuseDSL federated museum interoperability framework.

A node retains authority over its local collection database and determines which fields or records are exposed for interoperable access. The framework introduces a semantic interaction layer above existing publication standards rather than replacing them. Its architecture contains three principal layers: museum nodes, a federated interaction layer, and an agent-oriented processing layer.

1. Museum Nodes and Institutional Sovereignty

Each node represents a museum or collection system that publishes selected structured metadata through a web-accessible endpoint. The architecture allows institutions to retain local ownership and avoids mandatory replication of complete inventories into a central database. A museum can participate using JSON-LD, LIDO-derived metadata, or another structured representation, provided that the exposed fields can be mapped to the operational representation used by the matching workflow. This separation between local storage and interoperable interaction is important for institutional autonomy, access control, and gradual adoption.

2. Museum Interoperability Protocol

The Museum Interoperability Protocol (MIP) defines the logical interactions required by the framework. In the current prototype, MIP is implemented as lightweight API-based operations rather than as a final internet standard. Its core capabilities are node discovery, query forwarding, candidate metadata retrieval, link proposal, validation state exchange, and relationship lifecycle management. A source node can submit a query object to participating nodes, receive candidate records, and later send a proposed relationship for review without copying the complete remote inventory.

3. M-CORE and Agent-Oriented Processing

Heterogeneous candidate records are mapped into M-CORE, a shared internal representation that contains the fields required by the matching process. Depending on the collection type, these fields may include title, category, subtype, material, provenance, locality, taxonomic descriptors, geological age, and auxiliary metadata. The goal is operational normalization rather than universal ontology replacement, and mapping is performed only for the fields needed by the active MuseDSL rule.

The processing layer separates four responsibilities. The Discovery Agent retrieves candidates from federated nodes and external authorities. The Normalization Agent harmonizes values and maps records into M-CORE. The Matching Agent executes MuseDSL rules and calculates evidence scores. The explanation and governance functions produce traceable decision artifacts and route uncertain cases to curators. This modular separation improves maintainability and makes it possible to inspect which component produced each transformation or decision.

4. MuseDSL Rule Model

MuseDSL expresses curatorial matching logic as explicit, reusable rules. A rule specifies the scope of the comparison, the metadata fields to be compared, field-specific similarity functions, evidence weights, optional conditions, decision thresholds, and the relationship type that may be proposed. For example, a mineral rule can prioritize the normalized mineral name and the chemical formula, whereas a fossil rule can emphasize taxonomic name, subtype, geological age, and locality. This design makes the assumptions of the matching

process visible to domain experts and allows rules to evolve independently of the general software infrastructure.

$$S(q, c) = \sum w_i s_i, \qquad \sum w_i = 1 \tag{1}$$

In (1), q is the query object, c is a candidate object, s_i is a field-level similarity contribution, and w_i is the corresponding rule weight. The system converts the composite score into one of three states: ACCEPT, REVIEW, or REJECT. The REVIEW state forms a deliberate governance band for cases that are plausible but not strong enough for automatic acceptance.

IV. METHODOLOGY AND OPERATIONAL WORKFLOW

1. Research Design

The study follows a design-science-oriented approach in which the research artifact is developed to address a practical interoperability problem and is then evaluated through operational use. The objective is not to claim that a single similarity function solves all museum entity resolution problems. Instead, the study evaluates whether federated discovery, explicit domain rules, explanation generation, and curator-controlled validation can be integrated into a coherent workflow across heterogeneous web-published inventories.

2. Sequential Workflow

A query begins when an object is registered or selected within a participating museum. The Discovery Agent sends selected metadata to the federated search layer. Other nodes return candidate records and, where relevant, query an external authority source as well. The Normalization Agent maps the records into M-CORE, resolves common lexical variation, and prepares the fields required by the active rules. The Matching Agent then evaluates the candidates with MuseDSL. For each candidate, the system stores the activated rules, field-level similarity values, composite score, proposed relationship type, and decision state. ACCEPT candidates can be presented as high-confidence proposals, REVIEW candidates are routed to a curator, and REJECT candidates remain unlinked.

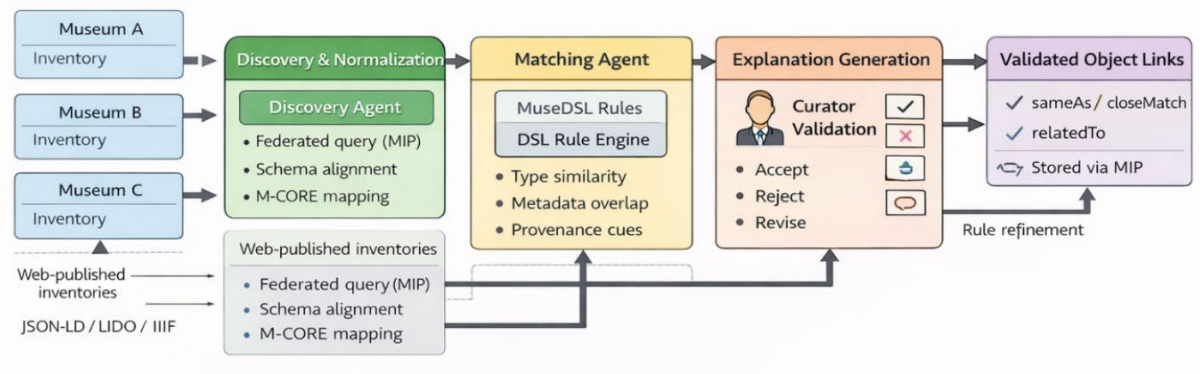


Figure 2. Operational workflow from federated discovery to curator-validated object links.

3. Explainability and Governance

An explanation artifact represents every candidate relationship. At minimum, the artifact contains the active MuseDSL rule, the compared fields, the field-level similarity values, the composite score, the decision state,

and the proposed link type. This trace allows a curator to understand not only what the system recommends but also why the recommendation was generated. The governance model treats a relationship as a lifecycle object rather than a one-time classification, so a proposed link can be reviewed, accepted, rejected, revised, or later withdrawn as museum knowledge changes.

4. External Authority Enrichment

The pilot uses Mindat as an auxiliary mineralogical authority source [14]. Mineral records can be submitted for name normalization, classification checking, synonym handling, and evidence enrichment. External authority evidence does not automatically determine the final cross-museum relationship; it is one source of evidence within the explanation chain. This distinction preserves the separation between external scientific references and institutional curatorial responsibility.

V. PROTOTYPE EVALUATION

1. Federated Pilot Environment

The prototype was evaluated in a web-based, federated pilot involving five museum inventory nodes, anonymized as Nodes A-E. Each node exposed selected metadata through API-based interfaces. No central index or pre-aggregated repository was created for the evaluation. A query object originating from one node could be sent to the other nodes, candidate metadata could be retrieved on demand, and the returned records could be evaluated through the common MuseDSL pipeline. The evaluation was designed around five practical questions: whether candidate objects could be discovered across nodes; whether collection-specific MuseDSL rules could be executed on heterogeneous records; whether an external authority could improve mineral name normalization; whether an auditable explanation could accompany every proposal; and whether uncertain results could be routed to human review rather than being forced into a binary decision.

2. Inventory and Task Pool

The evaluation inventory contained 1,000 web-published museum records distributed across five collection categories. The dataset is dominated by geological and natural history objects, making it suitable for testing rules that depend on mineral, rock, fossil, and biological descriptors. A batch of 100 source objects was used as an agent task pool. For each source object, the Discovery Agent queried the connected nodes, retrieved candidates, invoked authority matching for mineral-related records where applicable, and passed the results through normalization, matching, explanation, and decision routing.

Table 2. Distribution of object categories in the evaluation inventory

Category	Records	Percentage
Minerals	644	64.4%
Rocks	137	13.7%
Fossils	162	16.2%



Category	Records	Percentage
Biological specimens	39	3.9%
Other objects	18	1.8%
Total	1,000	100.0%

3. Federated Discovery and Rule Execution

The pilot confirmed the operational feasibility of the federated workflow. Candidate records could be retrieved from participating nodes without constructing a centralized collection database, and then normalized and evaluated using collection-specific rule families. Category gating prevented obviously incompatible categories from entering the same rule path, while M-CORE mapping enabled the comparison of differently structured fields through a shared operational representation. The architecture, therefore, supported interoperability as an interaction process rather than as a prior homogenization of global schemas. The strongest practical value of the rule approach was transparency. A curator can inspect whether a candidate was supported by name normalization, category compatibility, material similarity, taxonomic correspondence, geological age, locality, or another rule-defined feature. In contrast to a single opaque similarity score, the explanation artifact makes the computational path available for review. It enables revision of the rule when domain experts disagree with its behavior.

4. Mindat Authority Matching

Minerals formed the largest category, with 644 records. These records were submitted to the external authority matching process for name normalization and evidence enrichment. After normalization, the authority-matching rate exceeded 90% for the mineral subset, indicating that the majority of local mineral labels could be associated with recognized external entries. Unmatched or weakly matched records could, in turn, be treated as candidates for curator inspection. Possible causes of non-matches include spelling variants, synonyms, incomplete labels, local terminology, and records that require more detailed scientific review.

The authority result should not be interpreted as a final accuracy score for cross-museum object linking. It measures the success of associating local mineral labels with an external mineralogical authority after normalization. The broader object linking task also depends on the semantic relationship being sought and on additional evidence fields. Nevertheless, the result demonstrates the value of integrating external scientific authorities into an explainable rule-based workflow.

5. Explainability and Human Review

For the evaluated candidate relationships, the prototype generated rule-activation traces, field-level similarity decompositions, composite scores, proposed relationship types, and decision states. Cases that did not clearly satisfy the acceptance or rejection conditions could be routed to REVIEW. The review state is a central design feature because it prevents the system from presenting every computationally plausible link as an institutional fact. It also allows the framework to capture curator feedback and, in future implementations, to support controlled rule refinement. The current conference evaluation establishes feasibility rather than exhaustive comparative superiority. Large-scale baseline comparisons, ablation studies, threshold sensitivity analyses,



inter-rater agreement studies, and curator workload measurements are intentionally reserved for the expanded journal evaluation. This separation keeps the conference paper focused on the core artifact and the verified pilot observations, while leaving sufficient methodological space for a substantially extended archival publication.

VI. DISCUSSION

1. From Data Exchange to Governed Semantic Interaction

The main conceptual contribution of MuseDSL is to reposition museum interoperability as a process of forming governed relationships. Existing standards answer important questions about how cultural heritage information should be represented, exchanged, harvested, or displayed. MuseDSL addresses a different operational question: once several institutions expose compatible metadata, how can a new or existing object actively discover related objects, how can the evidence for a candidate relationship be expressed, and how can museum professionals remain responsible for the final decision? This distinction explains why the proposed framework is complementary rather than competitive. A museum node can continue to use CIDOC CRM, LIDO, IIIF, OAI-PMH, JSON-LD, or institution-specific databases. MuseDSL operates above these technologies as a rule-governed interaction layer. Its value depends on the availability of structured metadata, but it does not prescribe a single storage system or require participating museums to surrender control over their records.

2. Why a Domain-Specific Language Matters

Museum object linking is knowledge-intensive. The evidence that matters for two mineral records differs from that for two fossils or two archaeological artifacts. A general-purpose similarity function can be useful for retrieval, but it obscures domain assumptions when all fields are collapsed into a single implementation. MuseDSL exposes those assumptions as rules that can be read, reviewed, versioned, and adapted by domain communities. This is particularly important when different institutions use different acceptance thresholds or prioritize different evidence fields. The rule language also creates a practical bridge between software engineering and curatorial knowledge. Developers can implement reusable matching operators, while curators can reason about rule scope, field relevance, evidence weights, and relationship types. The resulting explanation trace provides a shared object for technical debugging and scholarly review.

3. Limitations

The pilot has several limitations. First, five nodes are sufficient to demonstrate distributed operation but not to establish national or international scalability; larger networks may introduce latency, availability, rate-limiting, duplicate-record, and multilingual metadata challenges. Second, the inventory is dominated by geological and natural history collections, so the current rule set cannot be assumed to apply directly to art-historical, ethnographic, archival, or archaeological domains. Third, the reported Mindat matching rate reflects the authority association after normalization and should not be conflated with end-to-end object link precision or recall. Fourth, the present conference paper does not report a complete quantitative comparison against alternative entity resolution models. These limitations define the next stage of evaluation. The extended study will compare title-only, multi-field lexical, authority-assisted, and full MuseDSL configurations; perform ablation studies of category gating, semantic normalization, authority enrichment, rule weighting, and the review band; examine threshold sensitivity; and evaluate the usefulness of explanations and museum

professionals' trust in them. Scalability tests will also measure response time, failure rate, throughput, candidate volume, and curator review workload under larger federated conditions.

VII. IMPLEMENTATION DETAILS AND REPRESENTATIVE RULE BEHAVIOUR

1. Federated Query and Relationship Lifecycle

A MuseDSL interaction begins with a source object that belongs to an initiating museum node. The source can be either a newly registered record or an existing object selected for cross-collection exploration. The initiating node prepares a compact query package that contains only the fields required by the active rule profile. This minimizes unnecessary data transfer and allows each institution to retain local control over fields irrelevant to the matching task. The query package identifies the source node, the object identifier, the requested collection scope, the active rule family, and the selected descriptive metadata. Participating nodes respond with candidate records rather than complete inventory exports. The Discovery Agent collects these responses and records the originating node for each candidate. This provenance information is preserved throughout the pipeline, because a relationship proposal is not merely a link between two metadata strings; it is a relationship between institutional records governed by different authorities. After normalization and matching, a proposal can be returned to the relevant target node, along with its evidence artifact and proposed relationship type. The lifecycle of a relationship is modeled through explicit states. A candidate may first exist as discovered and then become proposed after rule evaluation. High-confidence proposals enter the ACCEPT recommendation state; uncertain proposals enter REVIEW; and incompatible candidates are marked REJECT. A curator can validate, reject, or revise a proposal. A validated link can later be re-examined when new evidence appears, when an object record changes, or when a rule profile is updated. This lifecycle orientation is important because museum knowledge is not static: taxonomic classifications, provenance interpretations, object names, and institutional policies may all change over time.

2. Normalization and Field-Specific Similarity

MuseDSL separates normalization from matching. This design prevents every rule from re-implementing the same cleaning operations and makes the transformation process easier to inspect. Common normalization operations include case folding, whitespace trimming, removal or harmonization of diacritics, controlled synonym mapping, multilingual label alignment, and conversion of selected local fields into M-CORE. Domain-specific operations can also be attached to a field. For minerals, these operations may include mapping mineral names to synonyms and authority lookups. For fossils and biological specimens, taxonomic name normalization and rank-aware comparison may be used. For rocks, controlled vocabularies can represent rock class, texture, and genesis. Field-specific similarity is necessary because museum metadata fields have different semantics. A title may be compared using token overlap or normalized label equivalence, whereas a category field may require exact compatibility. Geological age can be compared using period mapping, locality using gazetteer-aware fuzzy matching, and numerical descriptors using tolerance functions. MuseDSL therefore treats similarity operators as reusable functions selected by rules rather than hard-coded into a single global scoring model. The composite score summarises the active evidence but does not replace the explanation. If a candidate achieves a score of 0.88, a curator should still be able to see whether that value was produced by strong taxonomic agreement, a shared material, a matching geological period, or several weaker signals. For this reason, the system stores the field-level contributions alongside the total. Optional fields are also distinguished from mandatory constraints: a missing locality, for example, should not necessarily invalidate a mineral species match, whereas a category mismatch may be sufficient to prevent a rule from executing at all.

3. Representative MuseDSL Rule Families

The prototype currently contains representative rule families for minerals, rocks, fossils, and biological specimens. The mineral rule family places high importance on the normalized mineral label and can use chemical formula, mineral group, and hardness as supporting evidence when those fields are available. The rock rule family emphasizes rock type and rock class, with texture and genesis as additional evidence. The fossil and biological rule family uses category compatibility, taxonomic name, specimen subtype, geological age, and locality. These examples demonstrate the language's central design principle: each collection domain can express its own evidence hierarchy while using the same federated execution and governance infrastructure. A rule also identifies the relationship that it is allowed to propose. An exact mineral species rule may propose a same-type or same-concept relationship, whereas a contextual rule may propose close-match or related-to. Separating the relationship type from the similarity score is important because two objects can be strongly similar without being identical, and a high score should not automatically imply semantic equivalence. The curator can therefore assess both the numerical evidence and the meaning of the proposed link. Rule versioning is another practical requirement. When a museum changes a threshold, adds a synonym, or modifies the relative importance of an evidence field, the resulting decisions should remain traceable to the rule version that produced them. The prototype records rule activations in the explanation log, which provides the basis for future versioned governance. In a larger deployment, institutions could maintain local rule profiles while also sharing community-approved rule libraries for common collection domains.

4. Representative Operational Scenario

Consider a museum that registers a fossil specimen with a scientific or local name, a specimen subtype, a geological period, and a discovery locality. After the record is saved, the museum can initiate a federated search. The source metadata is sent only to selected participating nodes. Candidate objects returned by other museums are normalized to M-CORE, and the fossil rule family evaluates taxonomic correspondence, subtype similarity, geological age, and locality. A candidate with strong taxonomic and chronological agreement may be proposed as a close match, whereas an object that shares only a broad category may be rejected. If the score falls within the review band, the curator receives the evidence rather than an automatic decision being made.

The same workflow can be applied in the opposite direction. A museum that adopts the same interoperability software can search other participating institutions when its own staff registers or reviews an object. Consequently, the architecture does not depend on a single central museum or a privileged aggregator. Each node can initiate discovery, and each institution can decide whether to expose metadata or accept proposed links. Over time, the result can become a distributed network of curator-validated relationships among museum objects. The prototype interface visualizes this process by allowing an initiating node and the active remote nodes to be selected, executing a federated query, and presenting the returned candidates with their source node, title, category, material, score, decision, and proposed link type. Separate evidence panels record the rule traces and field contributions. The interface is intentionally simple because the research contribution focuses on the interoperable workflow and the rule model rather than on a final production user experience.

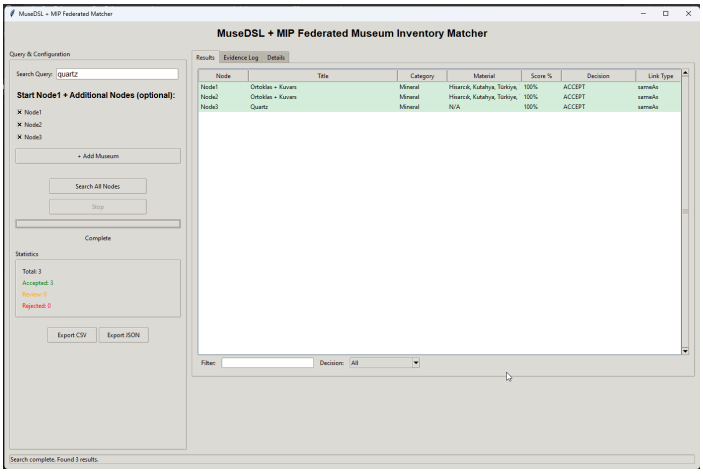


Figure 3. Prototype federated query interface showing candidate objects, decision states, and proposed link types.

5. Reproducibility, Security, and Deployment Considerations

The prototype and the representative rules are published as open-source software so that the architecture can be inspected and extended. Reproducibility in a federated setting, nevertheless, involves more than just source code. Live endpoints can change, external authority services may return updated records, and museum systems may apply access restrictions. For this reason, a robust reproducibility package should contain anonymized sample inventories, fixed API response snapshots, example rule files, exported explanation artifacts, and scripts that replay selected evaluation tasks without relying on live services.

Security and institutional control also require attention in future deployments. A production MIP implementation should define authentication, authorization, rate limiting, request logging, and provenance, along with clear policies for which fields may be exposed. Relationship proposals should be digitally attributable to their initiating node, and validation actions should record the responsible institution and the rule version. These features are not fully formalized in the current pilot, but they are compatible with the layered architecture, as communication, processing, and governance are separated.

Scalability will depend on query routing and candidate reduction. Broadcasting every object to every museum would be inefficient in a large network. Future implementations can use collection profiles, category-aware routing, lightweight node registries, caching of authority normalization, and staged retrieval in which broad filters are applied before expensive rule evaluation. Because MuseDSL rules expose their scope, they can also guide query planning by declaring which categories and fields are required before candidate data are requested. These deployment considerations reinforce the conference paper's role as the architectural and feasibility foundation for a broader research program. The short form establishes the language, protocol concept, agent workflow, governance logic, and the initial multi-node pilot. A full journal extension can then evaluate performance and generalisability using larger networks, stronger baselines, controlled ablations, multilingual collections, curator user studies, and formal protocol specifications, without simply repeating the same contribution.

6. Governance, Rule Maintenance, and Institutional Adoption

Institutional adoption requires more than successful retrieval. Museums need to know who is responsible for a proposed relationship, which version of the rule produced it, what evidence was used, and whether an authorized expert has reviewed the proposal. MuseDSL, therefore, treats the provenance of decisions as part of interoperability. A production deployment should retain the initiating node, the target node, the rule identifier, the rule version, the evaluation time, the explanation artifact, and the curator action. These records can support later auditing and can prevent a validated relationship from becoming detached from the reasoning that originally justified it. Rule maintenance should also be collaborative but controlled. A shared rule library can accelerate adoption across institutions, yet museums may legitimately apply different thresholds or local vocabularies. The framework can therefore distinguish community rules from local profiles. A community rule defines reusable domain logic, while an institutional profile can adjust synonyms, optional fields, thresholds, or review policies without changing the underlying protocol. This approach preserves interoperability at the level of interaction while allowing museums to retain scholarly and organizational autonomy. A further adoption challenge is the curator's workload. Fully automated linking may create an unacceptable number of false positives, whereas sending every candidate to human review can make the system impractical. The three-state decision model is intended to balance these risks: high-confidence proposals can be prioritized for rapid confirmation, clearly incompatible candidates can be filtered out, and only ambiguous cases enter the review queue. Future evaluations should therefore measure not only precision and recall, but also the number of review cases generated per object, the time required for expert validation, and the extent to which explanation artifacts reduce review effort. These measures will be essential for judging whether federated object linking is operationally sustainable in day-to-day museum practice.

VIII. CONCLUSION

This paper introduced MuseDSL, an explainable and federated framework for curator-governed museum object linking. The framework combines distributed discovery, an operational normalization layer, a domain-specific rule language, agent-oriented processing, explanation generation, and human-in-the-loop validation. Its purpose is to extend museum interoperability beyond static metadata exchange towards accountable semantic interaction among autonomous collection systems. A prototype was evaluated in a web-based pilot with five museum inventory nodes and a 1,000-record inventory. A 100-object task pool was processed through the agent workflow, and mineral records were also evaluated against Mindat as an external authority. The mineral authority matching process achieved a rate above 90% after normalization. The prototype also generated inspectable rule traces, field-level evidence, composite scores, and decision states for candidate object relationships. The results provide initial evidence that meaningful cross-institutional object discovery can be implemented without centralizing complete museum inventories. That explainability and curator review can be embedded as first-class architectural requirements. The open-source prototype and the rule examples are available at https://github.com/cyasar/MuseDSL. Future work will extend the current pilot with larger federated networks, broader collection domains, comparative baselines, ablation studies, threshold analysis, and structured curator-centered evaluation.

ACKNOWLEDGEMENT

The authors thank the institutional and technical contributors who supported the development and testing of the MuseDSL prototype, and acknowledge the value of the open cultural heritage and mineralogical data infrastructures that enable interoperable museum research.

REFERENCES

[1] M. Wooldridge, An Introduction to MultiAgent Systems, 2nd ed. Chichester, U.K.: Wiley, 2009.
[2] B. O'Neill and L. Stapleton, "Digital cultural heritage standards: From silo to semantic web," AI & Society, vol. 37, no. 3, pp. 891–903, 2022.
[3] M. Doerr, "The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata," AI Magazine, vol. 24, no. 3, pp. 75–92, 2003.
[4] R. Stein and O. Balandi, "Using LIDO for evolving object documentation into CIDOC CRM," Heritage, vol. 2, no. 1, pp. 1023–1031, 2019.
[5] S. Snydman, R. Sanderson, and T. Cramer, "The International Image Interoperability Framework (IIIF): A community and technology approach for web-based images," in Proc. Archiving Conf., 2015, pp. 16–21.
[6] C. Lagoze and H. Van de Sompel, "The Open Archives Initiative: Building a low-barrier interoperability framework," in Proc. 1st ACM/IEEE-CS Joint Conf. Digital Libraries (JCDL), 2001, pp. 54–62.
[7] H. Van de Sompel, M. L. Nelson, C. Lagoze, and S. Warner, "Resource harvesting within the OAI-PMH framework," D-Lib Magazine, vol. 10, no. 12, 2004.
[8] R. V. Guha, D. Brickley, and S. Macbeth, "Schema.org: Evolution of structured data on the web," Communications of the ACM, vol. 59, no. 2, pp. 44–51, 2016.
[9] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," Science Robotics, vol. 4, no. 37, art. no. eaay7120, 2019.
[10] M. Fowler, Domain-Specific Languages. Boston, MA, USA: Addison-Wesley, 2010.
[11] M. Casillo, M. De Santo, R. Mosca, and D. Santaniello, "An ontology-based chatbot to enhance experiential learning in a cultural heritage scenario," Frontiers in Artificial Intelligence, vol. 5, art. no. 808281, 2022.
[12] O.-M. Machidon, A. Tavčar, M. Gams, and M. Duguleană, "CulturalERICA: A conversational agent improving the exploration of European cultural heritage," Journal of Cultural Heritage, vol. 41, pp. 152–165, 2020.
[13] T. Gill, "Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model," First Monday, vol. 9, no. 5, 2004.
[14] Mindat.org, "Mindat.org – Mineral and locality database." [Online]. Available: https://www.mindat.org/ (accessed Jul. 2026).

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

The Effects of Meta-Heuristic Optimization Algorithms on the Performance of Machine Learning Models for Alzheimer's Diagnosis

Ömer Faruk Çakır
Gazi University
Computer Engineering
Ankara, Turkey
ofarukcakir98@gmail.com
ORCID: 0000-0001-6326-5557

Aysun Çoşkun
Gazi University
Computer Engineering
Ankara, Turkey
aysunc@gazi.edu.tr
ORCID: 0000-0002-8541-9497

Adem Tekerek
Gazi University
Computer Engineering
Ankara, Turkey
atekerek@gazi.edu.tr
ORCID: 0000-0002-0880-7955

Abstract

As part of this study, a machine learning-based decision support system was developed for the classification of Alzheimer's disease, and the effects of different hyperparameter optimization algorithms on model performance were examined. The Decision Tree (DT) and Random Forest (RRF) algorithms, selected as classification models, were optimized using Grey Wolf Optimization (GWO), Bayesian Optimization (BO), and a Genetic Algorithm (GA). While there are studies in the literature on artificial intelligence for disease diagnosis, research comparing multiple optimization algorithms on the same models remains limited. Experimental results have shown that hyperparameter optimization improves model performance. Compared to unoptimized base models, the Random Forest model achieved an increase of approximately 0.39% in Receiver Operating Characteristic Area Under the Curve (ROC-AUC). In comparison, the Decision Tree model achieved an increase of approximately 1.39%. While the highest ROC-AUC was achieved with the Random Forest + Genetic Algorithm combination, Bayesian Optimization proved to be the most time-efficient approach. In conclusion, the models' discriminative performance demonstrates the impact of hyperparameter optimization on medical decision support systems.

Keywords: alzheimer's disease, decision tree algorithm, random forest algorithm, grey wolf optimization algorithm, Bayesian optimization algorithm, genetic algorithm

I. INTRODUCTION

Artificial intelligence (AI) applications are becoming increasingly widespread in the healthcare sector, particularly for their advantages in high-risk clinical decision-making. However, for these systems to be used safely in clinical practice, they must adhere to the principles of Trustworthy AI, such as fairness, transparency, reliability, and accountability [1]. One of the areas where this requirement is most evident is the diagnosis of Alzheimer's disease, a progressive and irreversible neurodegenerative disorder. Alzheimer's is particularly prevalent among individuals aged 65 and older, and the number of patients worldwide is increasing each year [2], [3]. Since there is no definitive cure for the disease, early diagnosis is critical for the timely implementation of appropriate medical interventions and the effective management of its course [4], [5]. Deep learning and machine learning methods offer significant potential for the early diagnosis of Alzheimer's disease, particularly given their strong performance in analyzing Magnetic Resonance Imaging (MRI) data [3], [6]. However, the performance of these models depends largely on the proper selection of hyperparameters. Failure to properly optimize hyperparameters—such as the learning rate, number of iterations, activation functions, and regularisation parameters—can reduce model performance and lead to problems such as overfitting [7], [8]. Therefore, this study aims to maximize predictive performance by optimizing the hyperparameter spaces of

machine learning models using various metaheuristic optimization methods to predict Alzheimer's disease early and with high accuracy [7],[9].

This study aims to improve the diagnostic performance of Decision Tree and Random Forest models for diagnosing Alzheimer's disease. To this end, the models' hyperparameters were optimized using Grey Wolf Optimization (GWO), Bayesian Optimization (BO), and Genetic Algorithm (GA). The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) values obtained were compared to identify the most successful meta-heuristic optimization approach. Additionally, method runtime was analyzed to provide a comprehensive evaluation of accuracy and performance. The specific objectives for achieving this general goal are as follows;

1. What are the optimal hyperparameter values obtained for DT and RF models using GWO?
2. What are the optimal hyperparameter values obtained for DT and RF models using BO?
3. What are the optimal hyperparameter values obtained for DT and RF models using GA?
4. When comparing the performance of the GWO, BO, and GA optimization methods, which method produces more successful results, and is there a performance difference between the optimized and unoptimized models?
5. When examining the ROC curves of the models obtained after optimization, what is their classification performance?

II. RELATED STUDIES

A comprehensive literature review was conducted on artificial intelligence methods used in the diagnosis of Alzheimer's disease and the optimization techniques applied to them. It has been observed that artificial intelligence (AI) methods are increasingly being employed in the diagnosis of Alzheimer's disease. Mishra and Li [10] reported positive results for the application of AI in genetic analyses. Lale [11] and Özaltın [12] achieved high accuracy rates with the models they developed. At the same time, Bechir and Atasoy [13], Sukamto et al. [14], Zhang [15], and Bloch and Friedrich [8] reported that hyperparameter optimization improves model performance. El-Sappagh et al. [16] developed an explainable AI-based model for diagnosing Alzheimer's disease. Additionally, while Minesh [17] evaluated deep learning approaches, Vrahatis et al. [4] emphasized the significant potential of AI-assisted early-diagnosis studies. The literature indicates that artificial intelligence, machine learning, and deep learning methods are widely used in the early diagnosis and staging of Alzheimer's disease. Studies conducted using MRI images, clinical data, and biological markers demonstrate that these methods can achieve high accuracy. Research has shown that model performance depends not only on the algorithm used but also on hyperparameter optimization. While methods such as Grid Search, Bayesian Optimization, and Optuna can improve model performance, they can also reduce computational costs. However, challenges such as data insufficiency, class imbalance, interpretability, and high computational cost persist. Therefore, machine learning models supported by hyperparameter optimization represent an important area of research for the early, reliable prediction of Alzheimer's disease.

III. METHODS

The methodological workflow of the study begins with the collection and analysis of the dataset. Next, data preprocessing steps were applied, and training and test datasets were created. Class imbalance was addressed during this process, after which DT and RF models were built. Hyperparameter optimization was performed on these models, and their performance was evaluated using cross-validation. In the final stage, the results were compared to identify the best model.

1. Dataset Used

In this study, an Alzheimer's dataset comprising 2,149 patients, obtained from the "Kaggle" platform, was used to train machine learning models. The dataset was prepared to examine factors associated with the disease and to develop predictive models; however, it is synthetic. The dataset, which contains 35 numerical features, is categorized into groups such as patient ID, demographics, lifestyle, medical history, clinical measurements, cognitive assessments, and diagnostic information. No missing data was found in the dataset during the analysis.

In this study, the "PatientID" and "DoctorInCharge" attributes were removed from the patient dataset to ensure they did not influence the machine learning models, as they had no impact on disease diagnosis. Subsequently, 32 of the remaining 33 attributes were assigned to the set of independent variables, $X = \{x_1, x_2, \ldots, x_n\}$, while the remaining attribute was designated as the target variable for disease detection, $y \in \{0, 1\}$. The dataset was then split into 80% training data and 20% test data for the machine learning algorithms used in the study, and the "stratified sampling" approach was employed to preserve the class distributions. Finally, since the target class distribution in the dataset was not perfectly balanced and to prevent "majority class bias" in disease diagnosis, SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training data to generate new synthetic examples belonging to the minority class, thereby balancing the data distribution in the training set used for machine learning.

2. Machine Learning Models

In this study, the Decision Tree (DT) and Random Forest (RF) algorithms were used to classify Alzheimer's disease based on patient data. The reasons for choosing these algorithms include their interpretability, their widespread use in medical datasets, and the fact that they offer structures suitable for hyperparameter optimization.

1) *Decision tree (DT) algorithm:* In machine learning, the decision tree (DT) algorithm is one of the most popular supervised learning algorithms used today and can be applied to both classification and regression tasks [18],[19]. The decision tree algorithm divides data into different groups based on specific conditions. Each node in this structure represents a test for a specific variable, while each branch represents a possible outcome or decision of that test. The final decision at the end of the process is represented by a leaf node [18; 20]. In this study, the Gini impurity technique was used to split the data in the DT algorithm. The Gini index serves as an indicator of how "pure" the leaf nodes are for classification—that is, it indicates how mixed the training data assigned to each node is [21]. In this algorithm, the features that best separate the data and the threshold value are determined at each node. Therefore, the hyperparameters optimized in this algorithm are: "max_depth (maximum depth)," "min_samples_split (minimum number of samples per split)," and "min_samples_leaf (minimum number of samples per leaf node)." The formula for the Gini impurity technique used in the algorithm is: $Gini(D) = 1 - \sum_{i=1}^{C} p_i^2$, where the variable "pi" represents the class probability and the value "C" represents the total number of classes [21].

2) *Random forest (RF):* The basic operating principle of the RF algorithm is to combine a collection of random decision trees. The goal is to optimize a set of predictors (which are not necessarily optimal) by combining them, rather than optimizing a single predictor "at once" [22]. In this study, the RF algorithm—used for patient diagnosis—is an "ensemble learning" method that combines numerous decision trees. In this algorithm, each tree is created using the bootstrap sampling method. In the algorithm, the final class prediction is determined by majority voting, where "$\hat{y}$ = mode($T_1(x)$, $T_2(x)$, …, $Tn(x)$)" [23]. In this formula, "Ti(x)" represents the prediction of the i-th decision tree. The hyperparameters optimized in this algorithm are:

"n_estimators (number of estimators)", "max_depth (maximum depth)", "min_samples_split (minimum number of samples per split)", and "min_samples_leaf (minimum number of samples per leaf node)".

3. Hyperparameter Optimization

One of the most significant challenges in machine learning is the need to set initial parameter values for each learning algorithm before training begins; these parameters are referred to as hyperparameters. Hyperparameters can cause significant changes in the resulting model. Hyperparameters configure various aspects of a learning algorithm, such as the learning rate, kernel parameters, and network architecture [24]. In this study, the "Mean Cross-Validated ROC-AUC" was selected as the fitness function for hyperparameter optimization. The primary reason for choosing "ROC-AUC" in this comparison is that the study involves medical classification, and therefore, using the accuracy metric alone could be misleading. "ROC-AUC," on the other hand, is a threshold-independent metric that measures the balance between sensitivity and specificity. In this study, all optimization algorithms were evaluated using the same fitness function. The formula for the fitness function used in the study is: "$q^* = \arg max_{q \hat{I} W} f(q)$". Here, "θ" represents the hyperparameter values, "Ω" represents the search space, and "f(θ)" represents the objective function. The formula for "$f(q) = \frac{1}{K}\sum_{i=1}^{K} ROC_AUC_i$" and the "K" value in the formula is the "K-fold" number used in the "Stratified K-Fold Cross-Validation" method; in this study, this value was set to 5. The hyperparameter vectors required for the machine learning algorithms in this study are, in order:

- For DT: θ_DT=[max_depth, min_samples_split, min_samples_leaf],
- For RF: θ_RF=[n_estimators, max_depth, min_samples_split, min_samples_leaf]

4. Meta-heuristic Optimization Algorithms

Using the optimization algorithms employed in this study, the hyperparameter search performance of the machine learning algorithms DT and RF was evaluated. This made it possible to compare the effects of each optimization algorithm on classification performance.

1) Grey wolf optimization algorithm (GWO): GWO is an optimization algorithm based on swarm intelligence that mathematically models the hunting behavior of grey wolves [25]. In this algorithm, individuals within the swarm are classified as "alpha," "beta," "delta," and "omega." The best solution is considered to be the alpha wolf in the algorithm. The formulas for position updates [25] are: "$D = |C \cdot X_\alpha - X|$" and "$X(t+1) = X_\alpha - A \cdot D$", and the coefficient vectors here are "$A=2\sigma r_1 - \sigma$" and "$C = 2r_2$," where the values of "r_1" and "r_2" are random number vectors. The value of "σ" represents the control parameter that decreases throughout the iteration. In this study, the GWO algorithm was applied at each iteration as follows: the fitness value of each wolf was calculated; the wolf with the best fitness was selected as alpha; the other wolves were updated to converge toward alpha; new hyperparameter combinations were generated; and the ROC-AUC was recalculated. In this study, the number of wolves in the GWO algorithm was set to 10, and the number of iterations was set to 30.

2) Bayesian optimization algorithm (BO): In the BO algorithm, this approach is aimed at optimizing objective functions that take a long time to evaluate. It quantifies these functions using Gaussian process regression—a Bayesian machine learning technique—and then decides where to sample using a learning function defined by this model [26]. The general form of the formula used in the algorithm [27] is "$f(x) \sim GP(\mu(x), k(x, x'))$," where "GP" denotes the Gaussian process, μ(x) represents the mean function, and k(x, x′)

represents the covariance function. New hyperparameter points are selected using the acquisition function, and the formula [27] is "$x_{next} = \text{argmax}\alpha(x)$". In this study, the Bayesian Optimization algorithm was implemented using the "Optuna" library. Additionally, in this study, the "n_trials" value—the number of iterations—was set to 30. The implementation of the BO algorithm in this study was as follows: the first few trials were conducted randomly; the results obtained were modeled using a Gaussian or Parzen-based surrogate model; then, a balance between exploration and exploitation was achieved using the acquisition function; during this process, the most promising new parameters were selected; and finally, the ROC-AUC value was recalculated.

3) Genetic algorithm optimization olgorithm (GA): GA is a method for optimizing search algorithms for difficult problems based on the principle of genetic selection. The working principle of GA is as follows: similar to chromosome production in biology, it relies on genetic operations—such as selection, crossover, and mutation—applied to an initially random population to generate solutions across successive generations [28]. At the start of the algorithm, a population of randomly generated individuals is created and represented as "P(t) = {x_1, x_2, ..., x_n}". Each individual's fitness value is calculated using the objective function as follows: "$Fitness(x_i) = f(x_i)$". The progression of the GA algorithm involves the following steps in each generation (iteration): first, the selection process is performed to identify the best individuals; then, crossover is applied to recombine the parameters of the selected individuals; finally, in the mutation phase, random changes are made to the individuals. As a result of these operations, a new population was formed, and the fitness values were recalculated. In the GA algorithm, the population size was set to 10, and the number of generations (iterations) was set to 30; these parameters were used to search for the global optimum.

5. Performance Evaluation Criteria

In this study, results from models optimized with optimization algorithms were evaluated with the ROC-AUC metric. The ROC-AUC value provides a general assessment of model performance. This metric illustrates the relationship between the true positive rate (TPR) and the false positive rate (FPR). The calculations for TPR and FPR are as follows [29]: "$TPR = \frac{TP}{TP+FN}$" and "$FPR = \frac{FP}{FP+TN}$". In these calculations, the abbreviations used to evaluate the test and prediction performance of classification models denote the following states: TP for true positive, FN for false negative, FP for false positive, and TN for true negative.

IV. FINDINGS AND ANALYSIS

In this study, ROC-AUC was used as the primary metric to evaluate the performance of algorithms for diagnosing Alzheimer's disease. This metric was chosen because it more accurately reflects the model's ability to distinguish between disease classes. In addition, the model optimization times were compared. The results show that hyperparameter optimization directly affects the classification performance of both Decision Tree (DT) and Random Forest (RF) models.

1. Optimizing Hyperparameter Results with Meta-Heuristic Optimization Algorithms

1) Hyperparameter spaces obtained with the GWO algorithm: In this study, the GWO algorithm was used to identify the optimal hyperparameter spaces for both the DT and RF algorithms. The optimal hyperparameter space obtained for the DT model is as follows: "max_depth: 6.14269925," "min_samples_split: 6.23506442," and "min_samples_leaf: 5.8006925." The decimal values in this hyperparameter space were converted to integers for use in the algorithm. Upon examining this hyperparameter space obtained for the DT model, it can be seen that the tree depth is kept limited, the minimum number of samples per split is increased, and more samples are retained at the leaf nodes. The optimal hyperparameter space obtained for the RF model is as

follows: “n_estimators: 281.45510932”, “max_depth: 15.82994271”, “min_samples_split: 2.53040449”, and “min_samples_leaf: 1.13892439”. These results indicate a high number of trees, a deeper tree structure, and low leaf node thresholds, suggesting that the model is oriented toward higher representational capacity. Since the RF uses an ensemble learning approach, the risk of overfitting is lower, enabling the selection of more aggressive parameters with the GWO optimizer.

2) *Hyperparameter spaces obtained using the Bayesian optimization (BO) algorithm:* In this study, the BO algorithm was used to identify the optimal hyperparameter spaces for both DT and RF. The optimal hyperparameter space obtained for the DT model is as follows: “max_depth: 5”, “min_samples_split: 17”, “min_samples_leaf: 9”, and “Criteria: entropy”. Looking at this hyperparameter space, the BO algorithm produced parameters that are more resistant to change; in particular, the model maintains a simpler structure with “low max_depth,” “high min_samples_split,” and “high min_samples_leaf” values. Selecting entropy as the criterion, along with the metaheuristic optimization algorithm, also indicates that an information-gain-based splitting strategy was preferred in the DT model. Thus, a more orderly and controlled hyperparameter search was achieved using the BO algorithm. The optimal hyperparameter space obtained for the RF model is as follows: “n_estimators: 282,” “max_depth: 24,” “min_samples_split: 14,” “min_samples_leaf: 1,” and “criterion: entropy.” An evaluation of this hyperparameter space reveals that a high number of trees were generated for the RF model, resulting in a very deep structure. This enables the model to learn complex patterns while also limiting unnecessary branching.

3) *Hyperparameter spaces obtained with genetic algorithms (GAs):* In this study, the GA algorithm was used to identify the optimal hyperparameter spaces for both the DT and RF algorithms. The optimal hyperparameter space obtained for the DT model is as follows: “max_depth: 7,” “min_samples_split: 6,” and “min_samples_leaf: 8”—that is, a Decision Tree structure with medium complexity was established using the GA optimization algorithm. With this structure, the model maintains a balance between learning capacity and generalization ability. The optimal hyperparameter space obtained for the Random Forest (RF) model is as follows: “n_estimators: 222”, “max_depth: 28”, “min_samples_split: 3”, and “min_samples_leaf: 1”—that is, a Random Forest structure with high complexity was established using the GA optimization algorithm. Furthermore, because of the ensemble nature of the Random Forest, this complexity did not lead to significant overfitting.

2. Results Obtained with Optimization Algorithms

The results obtained from the optimization algorithms used in the study are given in Table 1. ROC-AUC sorts the results in Table 1.

Table1. Optimization algorithm results

Model	Meta Heuristic Optimization Algorithms	ROC-AUC	Time (s)
Random Forest (RF)	GA	0.942777	183.56
Decision Tree (DT)	Bayesian	0.942683	2.33
Random Forest (RF)	Bayesian	0.942517	27.64
Random Forest (RF)	GWO	0.942091	286.69
Decision Tree (DT)	GA	0.942067	18.29
Decision Tree (DT)	GWO	0.941239	27.69
Random Forest (RF)	-	0.939000	4.40

Decision Tree (DT)	-	0.929800	6.38

An examination of Table 1 reveals that machine learning models with an optimizer achieved better results than those without one. It is observed that the ROC-AUC scores of machine learning models that use an optimizer are very similar, hovering around 0.94. This result demonstrates that the optimization algorithms applied to the machine learning models can successfully classify the dataset. Additionally, the best ROC-AUC was obtained with the RF model using the GA algorithm. When comparing the results of the RF model without the Meta-Heuristic optimization algorithm to the best result from the RF model using an optimization algorithm, the model's success rate increased by approximately 0.39% with the optimization algorithm. Similarly, when examining the results using the DT algorithm, the model's success rate increased by approximately 1.39% with the optimization algorithm. This demonstrates that optimizing the hyperparameter spaces of machine learning models using optimization algorithms can yield better results.

In models using metaheuristic optimization algorithms, computation times generally increased compared to models without optimization. In particular, it was determined that the RF model with BO performed faster and produced ROC-AUC values close to those of the RF model obtained with GA. This indicates that BO can deliver high performance with lower computational cost. It was observed that the GWO and GA algorithms, due to their swarm- or population-based structures, require longer computation times.

3. ROC Curve Analysis

Figure 1 shows the ROC curve plots for all machine learning models to which meta-heuristic optimization algorithms were applied. Looking at this figure, the curves for all models are positioned quite close to the upper-left corner. This indicates that the models produced a high true positive rate and a low false positive rate.

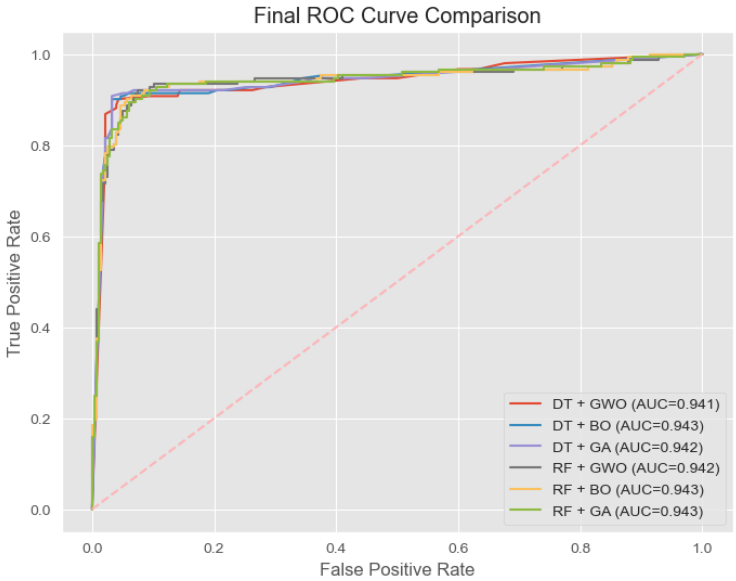


Figure 1. ROC curve for all algorithms.

Furthermore, as shown in the ROC curve, there are no significant differences in performance among the models. The main reasons for this are that the dataset used contains strong discriminative features and that all meta-heuristic optimization algorithms were able to reach suitable hyperparameter regions.

V. CONCLUSION

In this study, a comprehensive review of the literature was conducted on artificial intelligence models used in the diagnosis of Alzheimer's disease and the metaheuristic optimization methods employed for these models; studies that included literature reviews were examined in detail. Within this scope, machine learning models—specifically Decision Trees (DT) and Random Forests (RF)—were used to classify Alzheimer's disease. Meta-heuristic optimization techniques were applied to these machine learning models. The optimization processes were carried out using the GWO, BO, and GA algorithms. To evaluate the performance of machine learning models trained with these optimization algorithms, the ROC-AUC metric was used as the primary criterion for comparison.

A review of the literature revealed that artificial intelligence methods are widely used in the diagnosis of Alzheimer's disease, particularly in recent studies. Some studies have also shown that meta-heuristic optimization methods have been integrated into these models. However, no studies were found that directly compared the performance effects of applying multiple optimization methods to the same machine learning model.

The findings indicate that hyperparameter optimization directly affects model performance. When models without metaheuristic optimization were compared with optimized models, the ROC-AUC increased by approximately 0.39% for the RF model and 1.39% for the DT model. This increase was particularly pronounced in the DT model, indicating that hyperparameter optimization is more effective in simple models prone to overfitting. ROC-AUC values of approximately 0.94 were achieved across all methods, indicating high classification accuracy. The best results in DT were obtained with BO, while in RF they were achieved with GA; the RF+GA combination yielded the highest ROC-AUC.

In contrast, BO was observed to require less optimization time and to provide a more efficient search. The GWO algorithm, on the other hand, exhibited more stable convergence behavior. While the ROC curves revealed that all models demonstrated strong discrimination, with high TPR and low FPR, the RF model produced more balanced results than DT due to its ensemble structure.

A literature review shows that hyperparameter optimization improves model performance in studies aimed at predicting Alzheimer's disease. The findings of this study also support this approach. Throughout the process, the low computational cost of BO, GA's ability to search a wide solution space, and GWO's stable convergence behavior have highlighted the advantages of these different methods. In conclusion, metaheuristic and probabilistic optimization algorithms are effective in improving the performance of medical decision support systems.

REFERENCES

[1] El-Sappagh, S., Alonso-Moral, J. M., Abuhmed, T., Ali, F., & Bugarín-Diz, A. (2023). Trustworthy artificial intelligence in Alzheimer's disease: state of the art,

[2] Aditya Shastry, K., & Sanjay, H. A. (2024). Artificial Intelligence Techniques for the Effective Diagnosis of Alzheimer's Disease: A Review. Multimedia Tools and Applications, 83(13), 40057-40092.

[3] Zhang, W., Li, Y., Ren, W., & Liu, B. (2023). Artificial intelligence technology in Alzheimer's disease research. Intractable & Rare Diseases Research, 12(4), 208-212.
[4] Vrahatis, A. G., Skolariki, K., Krokidis, M. G., Lazaros, K., Exarchos, T. P., & Vlamos, P. (2023). Revolutionizing the early detection of Alzheimer's disease through non-invasive biomarkers: the role of artificial intelligence and deep learning. Sensors, 23(9), 4184.
[5] Elmotelb, A. S., Sherif, F. F., Abohamama, A. S., Fakhr, M., & Abdelatif, A. M. (2025). A novel deep learning technique for multi-classifying Alzheimer's disease: hyperparameter optimization—Frontiers in Artificial Intelligence, 8, 1558725.
[6] Yao, Z., Wang, H., Yan, W., Wang, Z., Zhang, W., Wang, Z., & Zhang, G. (2023). Artificial intelligence-based diagnosis of Alzheimer's disease with brain MRI images—European Journal of Radiology, 165, 110934.
[7] Emeç, M., & Özcanhan, M. H. (2023). Makine öğrenmesi algoritmalarında hiperparametre belirleme. Mühendislikte Öncü Ve Çağdaş Çalışmalar, 71-98.
[8] Bloch, L., & Friedrich, C. M. (2020, December). Using Bayesian optimization to effectively tune random forest and XGBoost hyperparameters for early alzheimer's disease diagnosis. In International Conference on Wireless Mobile Communication and Healthcare (pp. 285-299). Cham: Springer International Publishing.
[9] Tanyıldızı, E., & Demirtaş, F. (2019, November). Hiperparametre optimizasyonu (hyperparameter optimization). In 2019, the 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-5). IEEE.
[10] Mishra, R., & Li, B. (2020). The application of artificial intelligence in the genetic study of Alzheimer's disease. Aging and disease, 11(6), 1567.
[11] Lale, T. (2025). Alzheimer hastalığının erken tanısı için Optuna optimizasyonlu tabular transformer mimarisi. Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi, 28(4).
[12] Özaltın, Ö., Çobanbaş, S., Sırakaya, Y., & Güneş, Y. (2025). Alzheimer Hastalığının Manyetik Rezonans Görüntülerinden Hibrit Derin Öğrenme Yaklaşımı ile Otomatik Tespiti. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 37(1), 321-339.
[13] Bechir, M. K., & Atasoy, F. (2025). Enhancing Multi-Disease Prediction with Machine Learning: A Comparative Analysis and Hyperparameter Optimization Approach. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 1-1.
[14] Sukamto, Hadiyanto, & Kurnianingsih. (2023). KNN optimization using a grid search algorithm for the preeclampsia imbalance class. In E3S Web of Conferences (Vol. 448, p. 02057). EDP Sciences.
[15] Zhang, F., Petersen, M., Johnson, L., Hall, J., & O'Bryant, S. E. (2021). Accelerating hyperparameter tuning in machine learning for Alzheimer's disease with high-performance computing—Frontiers in Artificial Intelligence, 4, 798962.
[16] El-Sappagh, S., Alonso, J. M., Islam, S. R., Sultan, A. M., & Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Scientific reports, 11(1), 2660.
[17] Minesh, P. (2022). A review on the importance of artificial intelligence in Alzheimer's disease and its future outcomes for Alzheimer's disease. Research Journal of Pharmacology and Pharmacodynamics, 14(1), 13-22.
[18] Tavasoli, S. (2024, February 9). Top 10 Machine Learning Algorithms For Beginners: Supervised, and More. Simplilearn. https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article
[19] Reilly, J. (2023, November 24). Top 15 Machine Learning Algorithms: An In-Depth Guide. Akkio. https://www.akkio.com/post/top-15-machine-learning-algorithms-an-in-depth-guide
[20] Wakefield, K. (2024). A Guide To The Types Of Machine Learning Algorithms And Their Applications. SAS. https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html

[21] Kirk, M. (2017). Thoughtful Machine Learning with Python: A Test-Driven Approach. O'Reilly Media.
[22] Genuer, R., & Poggi, J. M. (2020). Random forests. In Random forests with R (pp. 33-55). Cham: Springer International Publishing.
[23] Shraban, S. S., Hassan, M., Bhuiyan, T., & Hassan, M. M. (2025, August). Sociodemographic Determinants of Gasoline Vehicle Ownership in the United States: A Machine Learning Approach. In 2025, IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) (pp. 328-333). IEEE.
[24] Pınar, M., Okumuş, O., Turgut, U. O., Kalıpsız, O., & Aktaş, M. S. (2017). Büyük veri içeren öneri sistemleri için hiperparametre optimizasyonu. Ulusal Yazılım Mühendisliği Sempozyumu, 22-272.
[25] Koc, I., Baykan, O. K., & Babaoglu, I. (2018). Gri kurt optimizasyon algoritmasına dayanan çok seviyeli imge eşik seçimi. Politeknik Dergisi, 21(4), 841-847.
[26] Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811.
[27] Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), 26-40.
[28] Lambora, A., Gupta, K., & Chopra, K. (2019, February). Genetic algorithm: A literature review. In the 2019 International Conference on Machine Learning, Big Data, Cloud, and Parallel Computing (COMITCon) (pp. 380-384). IEEE.
[29] Brownlee, J. (2019). Probability for Machine Learning: Discover How To Harness Uncertainty With Python. Machine Learning Mastery.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

An NLP-Based Review and Taxonomy of Heart Murmur Detection Algorithms (PhysioNet 2022)

Onuralp Arslan
Department of Biomedical Technologies
Izmir Katip Celebi University
İzmir, Turkey
onuralparslan97@gmail.com
0000-0001-7512-0224

Yalçın İşler
Department of Electrical and Electronics Engineering
Alanya Alaaddin Keykubat University
Antalya, Turkey
islerya@yahoo.com
0000-0002-2150-4756

Abstract
This report presents a comprehensive analysis of 44 research papers from the George B. Moody PhysioNet Challenge 2022 (Computing in Cardiology), focusing on automated heart murmur detection and clinical outcome prediction from Phonocardiogram (PCG) recordings. We examine the CirCor DigiScope dataset's clinical correlations, demographic distributions, and preprocessing requirements. To map the methodology landscape, we introduce an NLP-based framework comprising keyword co-occurrence analysis, spatial distance network visualization, and semantic text embeddings with Sentence Transformers. We demonstrate that a four-cluster taxonomy (k=4) effectively categorizes published approaches into traditional feature engineering, end-to-end deep learning, sequence/ensemble models, and advanced representation/multi-task learning. This review provides a structured overview of state-of-the-art modeling strategies and dataset characteristics for pediatric heart sound classification.

Keywords: Heart Murmur Detection, Phonocardiogram (PCG), Clinical Outcome Prediction, Deep Learning, PhysioNet Challenge

I. INTRODUCTION

This paper analyzes 44 research papers presented at the Computing in Cardiology Conference for the 2022 George B. Moody PhysioNet Challenge. The primary objective is to evaluate automated strategies for Heart Murmur Detection and Clinical Outcome Prediction from Phonocardiogram (PCG) recordings. Rather than proposing a new classification algorithm, this paper introduces a systematic, NLP-empowered review methodology for reliably analyzing mass publications in this domain. Our approach creates an easily replicable framework to cross-check methodological trends across literature. Using this pipeline, we identify top-performing architectures—such as hybrid Parallel Hidden Semi-Markov Models [1], Hierarchical Multi-Scale CNNs [2], and multi-task branches [3]—and systematically categorize the literature into four distinct paradigms: traditional feature engineering [4], standard end-to-end deep learning [5], sequence-to-sequence modeling with ensembles [6], and advanced representation learning [7].

II. DATA SOURCE CIRCOR DIGISCOPE DATASET

The dataset utilized in this analysis originates from The George B. Moody PhysioNet Challenge 2022 [8], which is built upon the CirCor DigiScope Phonocardiogram Dataset. To acquire the dataset reproducibly, raw files were downloaded directly from PhysioNet using the standard PhysioNet command-line utility (`wget -r -N -c -np https://physionet.org/files/challenge-2022/1.0.0/`).

The dataset comprises 3,163 recordings from 942 pediatric patients (1,452 unique patients across 1,568 encounters). Because 116 patients attended both screening campaigns, GroupKFold cross-validation grouped by patient ID is required to prevent data leakage. Processing this raw dataset requires aggressive denoising due to noisy clinical environments. For instance, to eliminate artificial interference from stethoscope placement and removal, Team MetaHeart discarded the first 5 seconds and last 3 seconds of every recording [9]. Similarly, Team IITH applied Butterworth bandpass filters (0.1 Hz to 300 Hz) to strip baseline wander and high-frequency ambient noise before feature extraction [10].

1. Signal Specifications

All recordings are sampled at 4,000 Hz, ensuring consistency and eliminating the need for resampling. However, sample durations range from 5.15 to 64.51 seconds, with most recordings lasting 20 or 30 seconds (Figure 1). A major protocol shift occurred between campaigns: 2014 recordings averaged 29.7 s, whereas 2015 recordings averaged 19.9 s. This discrepancy must be accounted for in dataset splits to prevent duration-based shortcut learning (Figure 2). The average recording duration is consistent across primary auscultation sites in both years (AV: Aortic Valve, PV: Pulmonic Valve, TV: Tricuspid Valve, MV: Mitral Valve), with only 4 recordings labeled as 'Phc' denoting uncommon recording locations (Figure 3).

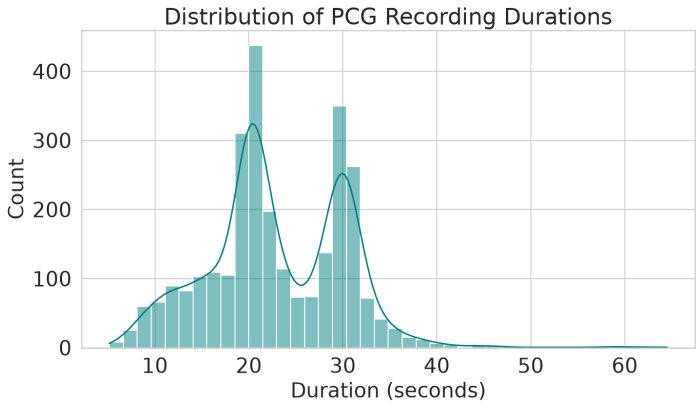


Figure 1. Signal Durations Distribution.

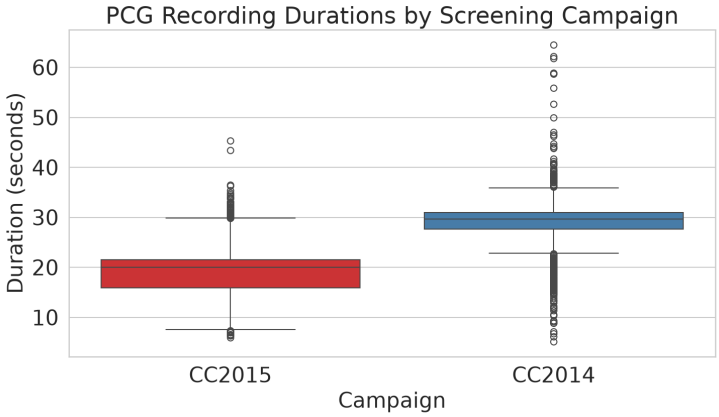


Figure 2. Signal Durations by Campaign.

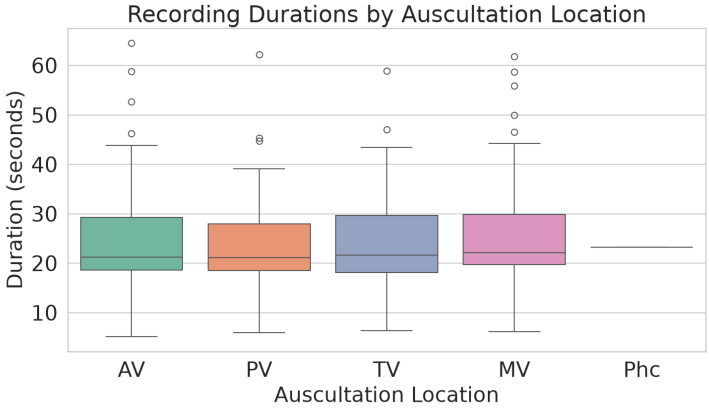


Figure 3. Signal Durations by Valve Location.

2. Demographic and Modalities

The dataset consists predominantly of school-aged children and is gender-balanced (Figure 4), with a 7.4% pregnancy rate (ages are anonymized for privacy). Approximately 11–12% of patients have missing physical metrics (Figure 5). Given this low missingness rate, discarding incomplete records presents a practical alternative to complex imputation.

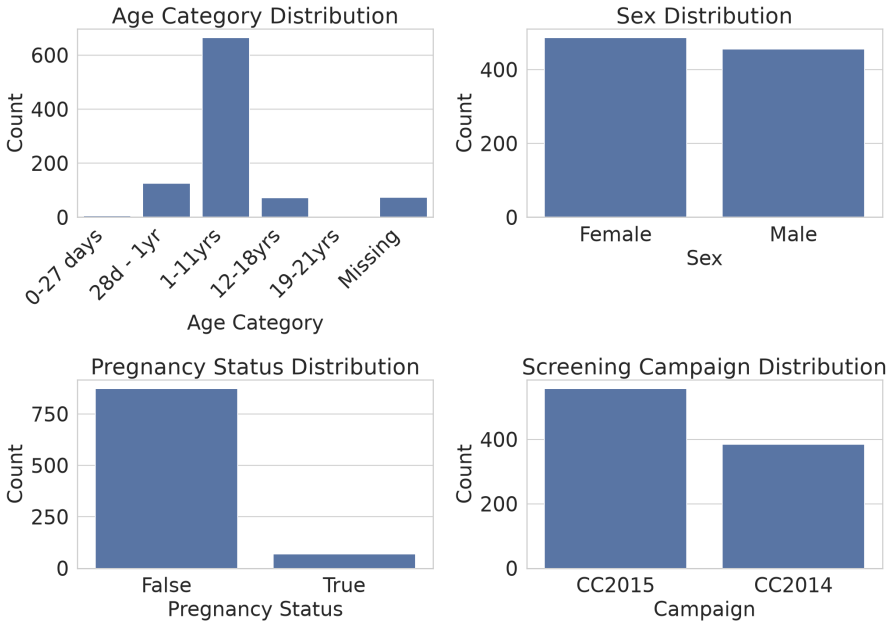


Figure 4. Patient Demographics.

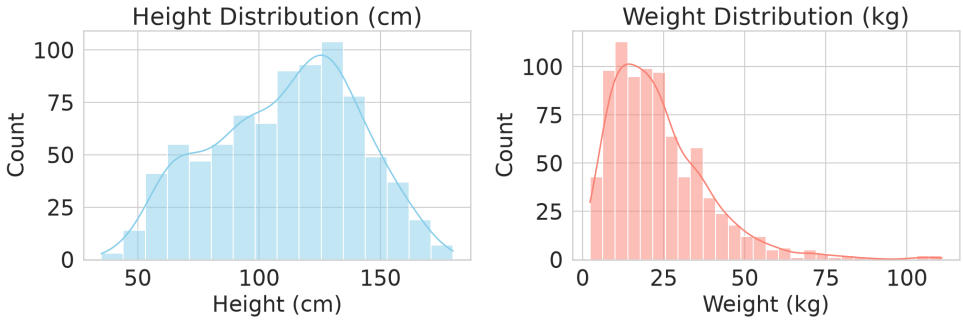


Figure 5. Height and Weight Distributions.

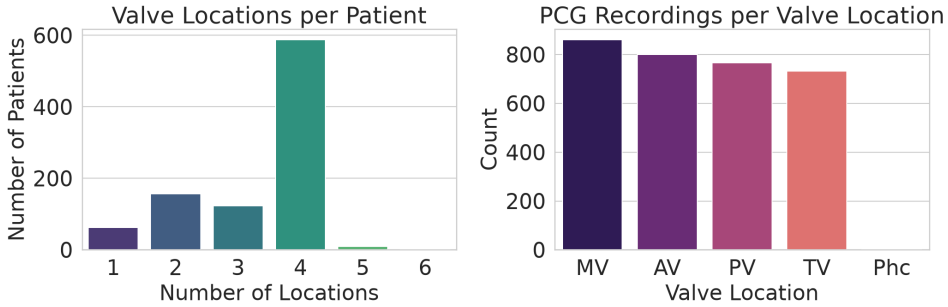


Figure 6. Modality Coverage.

Only 62.4% of patients had recordings across all four primary valves (Figure 6). This channel variability requires attention mechanisms or average pooling to dynamically aggregate features across varying input channels. To address the risk of overfitting due to missing leads, Team uke-cardio pioneered a novel CutMix augmentation strategy applied directly to auscultation locations. By artificially mixing PCG leads from different patients during training and generating soft-smoothed target labels, they forced their CNN to learn generalized acoustic features rather than patient-specific valve combinations [11].

3. Clinical Correlations and Targets

A severe class imbalance exists for murmur labels: 74% Absent vs. 19% Present (Figure 7). This imbalance necessitates weighted loss functions, such as Focal Loss. Conversely, clinical outcomes are relatively balanced, making a multi-task architecture with dual output heads an effective strategy. Team MetaHeart practically demonstrated this approach: to stabilize their 30-layer ResNet, they implemented a joint loss function that combined Center Loss (to contract intra-class variance in the deep feature space) with Class-Distinguished Focal Loss to heavily penalize false negatives on the minority "Present" class [9].

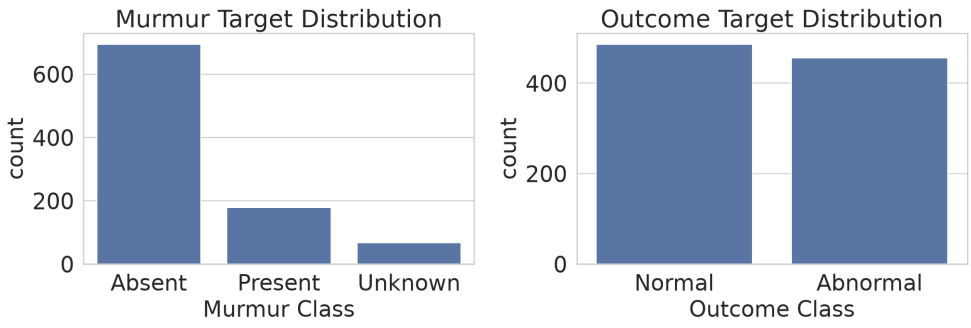


Figure 7. Class Imbalance.

Higher Levine grades are strongly associated with abnormal clinical outcomes (Figure 8), indicating that classifiers should leverage sound-intensity features to improve diagnostic accuracy. Because sound intensity and cardiac cycle phases (S1, systole, S2, diastole) are closely tied to clinical outcomes, top architectures utilize explicit segmentation as a regularization mechanism. For example, Team matLisboa trained a 567,000-parameter Transformer backbone with a dedicated auxiliary segmentation head, ensuring predictions were grounded in physiologically meaningful cardiac boundaries [12].

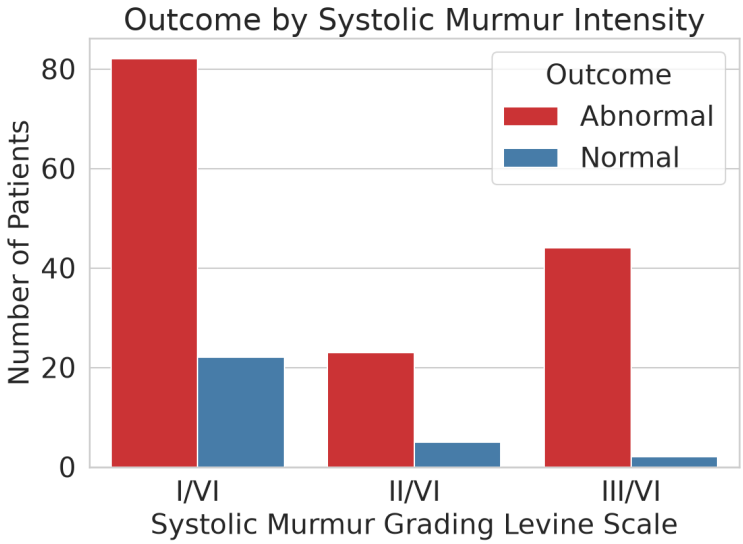


Figure 8. Murmur Grading vs. Outcome.

III. METHODOLOGY AND ANALYSIS

1. Data Acquisition and Preprocessing

To ensure complete reproducibility, all 44 challenge papers were programmatically downloaded from the PhysioNet 2022 open-source repository using Python's `requests` and `BeautifulSoup` libraries, and subsequently converted into structured Markdown text using `pymupdf4llm` for precise text-mining. This standardization allows for the robust application of NLP techniques across the entire corpus.

2. Keyword Co-occurrence

To reliably extract methodological trends, we used an automated NLP pipeline to parse the standardized Markdown abstracts and methodology sections of all 44 papers. The algorithm counts explicit mentions of predefined architectural keywords to generate a co-occurrence matrix (Figure 9). The matrix reveals that the strongest meaningful co-occurrence is between Segmentation and Cross-validation (as prominently seen in works such as [1] and [12]), indicating that explicit structural heart-cycle division, paired with rigorous validation, is the foundational standard for combating overfitting. Conversely, we identified rare, non-traditional combinations—such as pairing Wavelet Scattering directly with Support Vector Machines [4] or integrating Transformers with Ensemble methods [6]—which show minimal overlap with standard CNN baselines. Generic pairings, such as standard optimizers coupled with deep networks, were excluded from this analysis to isolate core architectural choices.

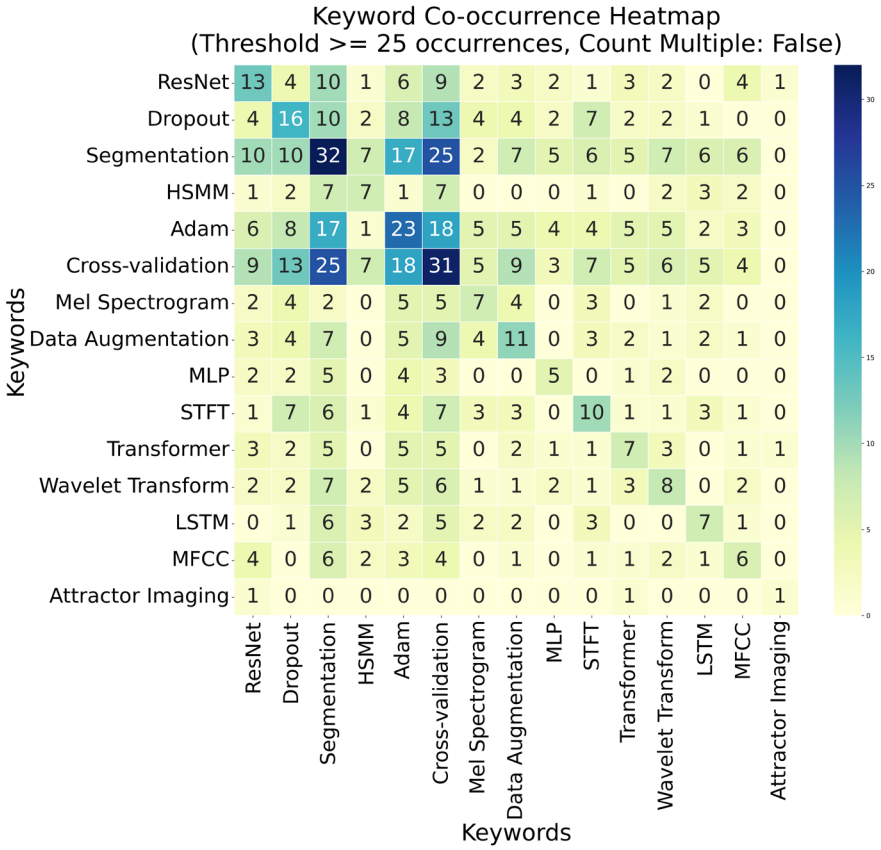


Figure 9. Keyword Co-occurrence Heatmap

3. Keyword Spatial Distance Networks

Beyond simple co-occurrence, the absolute character distance between methodology keywords was computed to build spatial semantic networks (Figure 10) at different thresholds. Visual inspection of the resulting network graph reveals a highly interconnected structure, particularly among classical machine learning methodologies. The thickest and darkest edges form a distinct, tight topological triangle between Principal Component Analysis (PCA), Support Vector Machines (SVM), and Random Forests. Furthermore, SVM maintains a prominent connection directly to Segmentation. This dense clustering suggests that classical machine learning approaches in the literature are treated as an inseparable package rather than dynamic, mixed methodologies. The consistently short character distances between these terms indicate a rigid standard workflow: audio segmentation, followed by dimensionality reduction via PCA, culminating in classification using standard algorithms. Conversely, an analysis of modern deep learning architectures, such as ResNet and Transformer, shows them to be positioned somewhat independently. They lack the strong connections observed in the classical feature-extraction clusters. This separation highlights a clear divide in literature: studies either rely on standardized classical pipelines or employ end-to-end deep learning networks, with very little hybridization between the two.

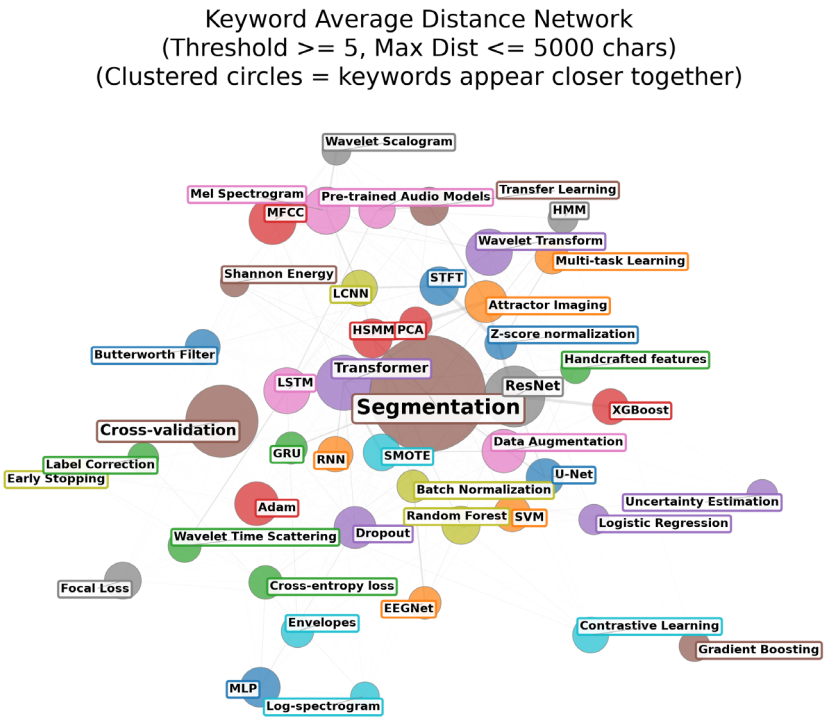


Figure 10. Keyword Distance Network

It must be acknowledged that textual proximity within a document is a relatively weak proxy for semantic relatedness when compared to high-dimensional contextual embeddings. While it successfully captures rigid, boilerplate descriptions of classical workflows (where methods are often listed in immediate succession, e.g., "features were extracted and classified using SVM"), it can fail to capture complex structural relationships that span multiple paragraphs. For this reason, true semantic categorization is performed using transformer-based models.

4. Semantic Embedding Clustering

To categorize publications semantically beyond simple keyword matching, we generated 384-dimensional text embeddings using a pre-trained sentence transformer (all-MiniLM-L6-v2). We clustered them directly on this high-dimensional space via K-means.

IV. RESULTS AND DISCUSSION

Following semantic embedding clustering, through heuristic trial-and-error and evaluation of the semantic coherence of the resulting groupings, we observed that k=2 is overly broad (grouping basic CNNs [5] with complex self-supervised architectures [7]), while k=5 results in severe over-segmentation due to trivial stylistic variations. Therefore, k=4 effectively segregates the core architectural differences. However, we acknowledge that this study currently lacks solid quantitative internal validation (such as computing silhouette scores or assessing stability across random seeds) to establish k=4 definitively. Formalizing a quantitative method to

determine optimal k alongside semantic embeddings remains a necessary focus for future work. The analysis concludes that k=4 effectively categorizes the solutions into four distinct paradigms. Cluster 1 represents traditional feature engineering, relying heavily on handcrafted heuristics such as wavelet time scattering, paired with classical classifiers, as in the works of Cornely et al. [4] and Imran et al. [13]. Cluster 2 groups supervised end-to-end deep learning baselines, where standard 1D/2D CNNs and ResNets process Mel-spectrograms without complex auxiliary objectives, as demonstrated by Peng et al. [5] and Lu et al. [14]. Cluster 3 captures sequence-to-sequence modeling and ensemble fusion, heavily featuring Transformer and LSTM networks to handle label noise and the unknown murmur class, as seen in studies by Alkhodari et al. [6] and Gemke et al. [15]. The most structurally complex approaches fall under Cluster 4 (advanced representation of learning and multi-task learning). Due to dataset class imbalance and variable recording lengths, these works move beyond simple supervised learning. Instead, they incorporate self-supervised pre-training [7] and multi-task optimization (using heart sound segmentation as an auxiliary task) to prevent overfitting. Additionally, these approaches target the challenge's asymmetric, cost-based scoring metric, which penalizes false negatives much more heavily than false positives. To optimize this metric directly, Cluster 4 architectures utilize custom loss functions—often paired with hybrid signal processing frontends or state predictions, as seen in McDonald et al. [1] and Rohr et al. [16].

V. CONCLUSION

In summary, this NLP-driven analysis provides a structured taxonomy of the PhysioNet 2022 Challenge, mapping the evolution from traditional feature engineering to advanced representation learning. By making the Python-based extraction and parsing pipeline transparent, this easily replicable methodology offers a clear view into the current state-of-the-art modeling strategies for automated heart sound classification. While our current taxonomic division into four paradigms aligns with the field's logical progression, we acknowledge that a heuristic evaluation of the embedding space initially guided it. Future iterations of this pipeline will employ rigorous quantitative internal validation to confirm the optimal number of paradigms mathematically.

REFERENCES

[1]. Andrew McDonald, Mark JF Gales, and Anurag Agarwal. Detection of heart murmurs in phonocardiograms with parallel hidden semi-Markov models. In 2022, Computing in Cardiology (CinC). IEEE, 2022.
[2]. Yujia Xu, Xinqi Bao, Hak-Keung Lam, et al. Hierarchical multi-scale convolutional network for murmurs detection on PCG signals. In 2022, Computing in Cardiology (CinC). IEEE, 2022.
[3]. Yale Chang, Luoluo Liu, and Corneliu Antonescu. Multi-task prediction of murmur and outcome from heart sound recordings. In 2022, Computing in Cardiology (CinC). IEEE, 2022.
[4]. Adrian K Cornely and Grace M Mirsky. Heart murmur detection using wavelet time scattering and support vector machines. In 2022, Computing in Cardiology (CinC). IEEE, 2022.
[5]. Guohui Peng, Haitao Zou, and Jin Wang. Classification of phonocardiograms using a residual convolutional neural network and an MLP. In 2022, Computing in Cardiology (CinC). IEEE, 2022.
[6]. Mohanad Alkhodari, Syafiq Kamarul Azman, Leontios J. Hadjileontiadis, and Ahsan H. Khandoker. Ensemble transformer-based neural networks detect heart murmur in phonocardiogram recordings. In 2022, Computing in Cardiology (CinC). IEEE, 2022.
[7]. Aristotelis Ballas, Vasileios Papapanagiotou, Anastasios Delopoulos, and Christos Diou. Listen2yourheart: A self-supervised approach for detecting murmur in heart-beat sounds. In 2022, Computing in Cardiology (CinC). IEEE, 2022.
[8]. M. A. Reyna, Y. Kiarashi, A. Elola, J. Oliveira, F. Renna, A. Gu, E. A. Perez Alday, N. Sadr, S. Mattos,

M. T. Coimbra, R. Sameni, A. B. Rad, Z. Koscova, and G. D. Clifford. "Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022" (version 1.0.0). PhysioNet (2023). https://doi.org/10.13026/t49p-5v35

[9]. Pan Xia, Yicheng Yao, Changyu Liu, Hao Zhang, Lirui Xu, Yuqi Wang, Lidong Du, Yusi Zhu, and Zhen Fang. Heart murmur detection from phonocardiogram based on a residual neural network with classes distinguished by focal loss. In 2022, Computing in Cardiology (CinC). IEEE, 2022.

[10]. Vijay Vignesh Venkataramani, Akshit Garg, and U Deva Priyakumar. Modified variable kernel length resnets for heart murmur detection and clinical outcome prediction using phonocardiogram recordings. In 2022, Computing in Cardiology (CinC). IEEE, 2022.

[11]. Marius S Knorr and Jan P Bremer. Using mel-spectrograms and 2d-cnns to detect murmurs in variable-length phonocardiograms. In 2022, Computing in Cardiology (CinC). IEEE, 2022.

[12]. João L Costa, Paula Couto, and Rui Rodrigues. Multitasking and transfer learning for cardiac abnormality detection in heart sounds. In 2022, Computing in Cardiology (CinC). IEEE, 2022.

[13]. Z. Imran, E. Grooby, et al. "A Fusion of Handcrafted Feature-Based and Deep Learning Classifiers for Heart Murmur Detection." In 2022, Computing in Cardiology (CinC). IEEE, 2022.

[14]. H. Lu, J. B. Yip, T. Steigleder, S. Grießhammer, et al. "A Lightweight Robust Approach for Automatic Heart Murmurs and Clinical Outcomes Classification from Phonocardiogram Recordings." In 2022, Computing in Cardiology (CinC). IEEE, 2022.

[15]. P. Gemke, N. Spicher, T. Kacprowski. "An LSTM-based Listener for Early Detection of Heart Disease." In 2022, Computing in Cardiology (CinC). IEEE, 2022.

[16]. M. Rohr, B. Müller, S. Dill, G. Güney, C. H. Antink. "Two-Stage Multitask-Learner for PCG Murmur Location Detection." In 2022, Computing in Cardiology (CinC). IEEE, 2022.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Enabling Clinical-Grade Left Ventricular Hypertrophy Detection in Wearable Devices via Deep Learning

Arda SARPAY
Institute of Natural and Applied Sciences
Izmir Katip Celebi University
Izmir, Turkey
arda_sarpay@hotmail.com
0000-0003-3143-6731

Yalcin ISLER
Dept of Electrical and Electronics Engineering
Alanya Alaaddin Keykubat University
Antalya, Turkey
islerya@yahoo.com
0000-0002-2150-4756

Abstract
Cardiovascular diseases remain one of the leading causes of mortality worldwide. Electrocardiography (ECG) is a widely used noninvasive diagnostic tool for detecting both electrical and structural cardiac abnormalities. While standard 12-lead ECG systems provide comprehensive clinical information, they are not suitable for continuous daily monitoring. Recent wearable devices such as smartwatches and portable ECG patches enable long-term ECG acquisition but typically provide only one to three leads, resulting in substantial information loss. This study proposes a deep learning framework for detecting left ventricular hypertrophy (LVH) and related cardiac pathologies using reduced-lead ECG signals. Unlike existing approaches that primarily focus on arrhythmia detection or explicit lead reconstruction, the proposed framework aims to transfer clinically meaningful representations from 12-lead ECG recordings to wearable reduced-lead configurations through a teacher–student knowledge distillation strategy. The PhysioNet/Computing in Cardiology Challenge 2021 dataset will be utilized for model development and evaluation. The proposed approach seeks to bridge the gap between clinical-grade ECG interpretation and wearable healthcare technologies, enabling early detection of structural heart abnormalities in everyday settings.

Keywords: Reduced-Lead ECG, Knowledge Distillation, Wearable Health, Left Ventricular Hypertrophy, Deep Learning, Hybrid Neural Networks

I. INTRODUCTION

Cardiovascular diseases constitute a major global health burden and account for millions of deaths annually. Early diagnosis of cardiac abnormalities is critical for reducing morbidity and mortality [1]. Electrocardiography remains one of the most accessible and cost-effective diagnostic modalities [2]. Conventional 12-lead ECG systems provide multiple spatial views of cardiac electrical activity and are routinely used for diagnosing arrhythmias, myocardial ischemia, conduction disorders, and structural abnormalities [2]. In contrast, wearable ECG devices have recently gained widespread adoption due to their ability to provide continuous monitoring outside clinical environments [3]. However, these devices are generally limited to one to three ECG leads, which significantly reduces the spatial information available for diagnosis [4]. Current wearable ECG research largely focuses on rhythm abnormalities such as atrial fibrillation. Structural cardiac abnormalities, particularly left ventricular hypertrophy (LVH), remain comparatively underexplored despite their clinical importance. This work investigates whether clinically meaningful information associated with LVH can be extracted from reduced-lead ECG recordings using modern deep learning techniques [1],[4].

II. BACKGROUND

LVH refers to abnormal thickening of the left ventricular myocardium, commonly caused by chronic hypertension and pressure overload [1]. LVH is associated with increased risks of heart failure, arrhythmias, stroke, and sudden cardiac death [1]. Traditional ECG-based LVH diagnosis relies on voltage criteria such as the Sokolow-Lyon and Cornell indices [4]. Although these criteria are clinically established, they often exhibit limited sensitivity and substantial inter-patient variability [4].

III. RELATED WORK AND LIMITATIONS

The PhysioNet/Computing in Cardiology Challenge 2021, entitled "Will Two Do?", evaluated algorithm performance across multiple lead configurations, including 2-, 3-, 4-, 6-, and 12-lead ECGs [5]. Top-performing solutions predominantly employed ensemble Convolutional Neural Network (CNN) architectures, attention modules, data augmentation strategies, and class-balancing techniques [6]. The PhysioNet Challenge 2021 provided a massive dataset of over 120,000 multi-lead ECGs to evaluate reduced-lead diagnostics [7]. A review of state-of-the-art methodologies reveals several critical limitations for wearable deployment [4],[5].

1. The Computational Burden of Ensembles

Top-performing architectures extensively utilized Deep Ensembles. Teams combined multiple deep Residual Convolutional Neural Networks (ResNets) or Squeeze-and-Excitation Wide ResNets to maximize the classification benchmark score [5],[6]. While highly accurate on cloud servers, executing deep networks in parallel requires significant memory and computational power, rendering them unsuitable for continuous execution on Edge IoT wearable microcontrollers [8].

2. Explicit Generative Reconstruction

To compensate for missing channels, some studies employ Generative models, Generative Adversarial Networks (GANs), or Autoencoders for medical time-series generation and signal reconstruction, explicitly reconstructing the missing 11 leads from a single lead before running a classifier [9]. In the medical domain, this introduces a severe risk of "hallucination"—the generative model might synthesize a healthy waveform over an actual ischemic event, potentially leading to clinically misleading predictions [1],[4].

3. Our Algorithmic Shift

Our approach explicitly rejects signal reconstruction and heavy ensembles [4]. Instead, we focus on representation learning. We hypothesize that the subtle morphological footprint of LVH is implicitly present in Lead I, albeit imperceptible to human cardiologists. Through Knowledge Distillation (KD), we guide a lightweight student network to map these subtle temporal features into a robust spatial latent space [10],[11]. Recent studies have highlighted the potential of KD in single-lead monitoring systems [12], establishing a strong baseline for expanding its application to structural pathologies.

IV. PROPOSED METHODOLOGY

The proposed framework is based on teacher–student representation learning. A conceptual overview of the proposed framework is shown in Figure 1. During training, a teacher network receives complete 12-lead ECG recordings and learns rich latent clinical representations from the full spatial information available in the signals. A student network receives only reduced-lead ECG signals (1–3 leads) and is trained to imitate the teacher's internal representations through knowledge distillation [10]. Unlike explicit lead reconstruction approaches, the

primary objective is not to regenerate missing ECG channels but rather to preserve diagnostically relevant information required for LVH detection [4],[9].

The overall pipeline consists of:

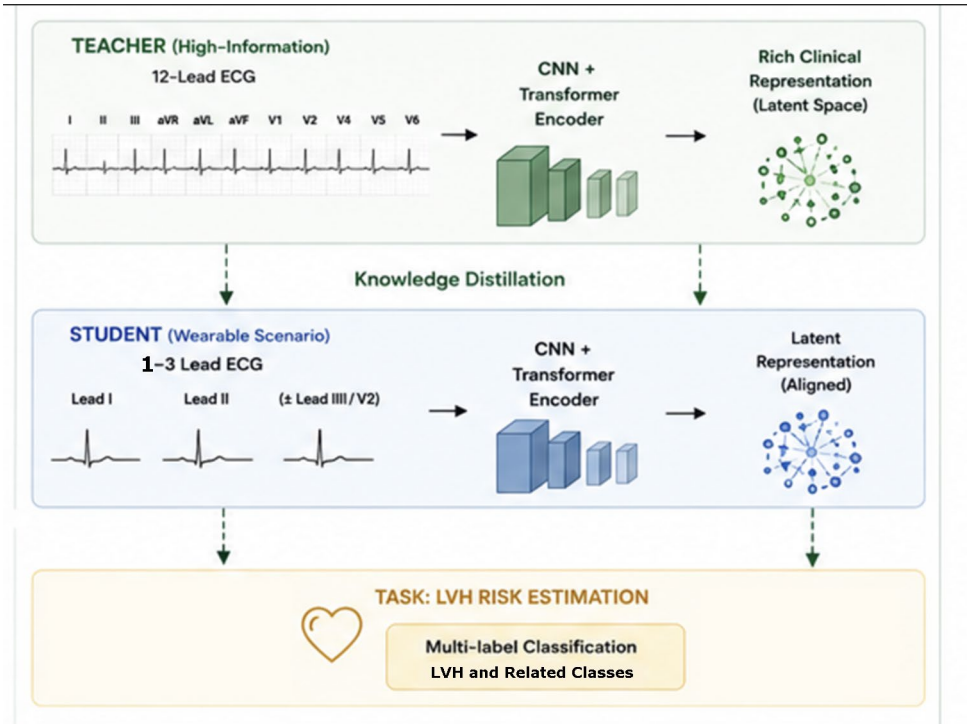


Figure 1. Overview of the proposed teacher–student knowledge distillation framework

V. DATASET AND EXPERIMENTAL PLAN

The study will utilize the PhysioNet/CinC 2021 Challenge dataset [5],[7]. The dataset contains approximately 121,144 ECG recordings collected from multiple institutions across different countries and patient populations [7]. The dataset incorporates recordings from multiple publicly available ECG databases, including CPSC, PTB-XL, INCART, Georgia, Chapman-Shaoxing, Ningbo, and UMich databases [7].

The training partition presents non-uniform signal properties that simulate deployment across heterogeneous wearable hardware:

- Signal Duration: The vast majority of records are exactly 10.0 seconds long, though the database exhibits a dynamic duration range from a minimum of 5.0 seconds to a maximum of 1,800.0 seconds (30-minute Holter monitoring) [7].
- Sampling Frequencies: The primary acquisition hardware operates at 500 Hz (99% of the records). Minor variations include 1,000 Hz (< 1%) and 257 Hz (only 74 records), necessitating a uniform resampling step during preprocessing [7].

- Lead Subsets: To evaluate wearable scenarios, the student network will be systematically tested by isolating discrete low-resource channel combinations, specifically 6-lead, 4-lead, 3-lead, and 2-lead (Leads I and II) setups [5].

Models will be validated using robust standard classification metrics, including Precision, Recall (Sensitivity), F1-Score, and the Area Under the Receiver Operating Characteristic (AUROC) curve. To ensure alignment with the international literature, the primary emphasis will be on optimizing reduced-lead LVH Recall and the official macro-weighted Challenge score [5]. Special emphasis will be placed on reduced-lead LVH detection performance and representation of transfer effectiveness.

VI. EXPECTED RESULTS AND DISCUSSION

We anticipate that the Student network, guided by KD, will significantly outperform baseline 1-lead models that lack Teacher supervision [12]. While the 1-lead Student may not completely match the absolute diagnostic precision of the 12-lead Teacher, the framework is expected to close the performance gap sufficiently to serve as a reliable probabilistic risk estimator [2], [10]. Furthermore, by avoiding Deep Ensembles, the resulting Student model will have a drastically reduced parameter footprint, making it well-suited for quantization and deployment on Edge AI microcontrollers embedded in smartwatches [8]. To prevent the model from operating as a black box, Explainable AI (XAI) techniques, such as Attention Saliency Maps, will be utilized to visually verify that the Student focuses on genuine physiological markers rather than background noise [13].

VII. CONCLUSION

This paper outlines a computer science solution to a critical medical problem: detecting structural cardiac diseases, such as Left Ventricular Hypertrophy (LVH), using comfortable, everyday wearable devices. By replacing computationally expensive ensembles and explicit generative reconstructions with a Teacher-Student Knowledge Distillation framework, the proposed approach aims to transfer deep clinical spatial representations to low-resource environments [10],[12]. Ultimately, this representation-level learning approach has the potential to bridge the gap between standard 12-lead clinical ECGs and reduced-lead wearables, eliminating the strict need for specialized clinical equipment [1]. It paves the way for a non-invasive, daily early-warning system that makes advanced, continuous cardiac screening widely accessible to everyone.

REFERENCES

[1]. N. Strodthoff, J. M. Lopez Alcaraz, and W. Haverkamp, "Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: An explorative study in emergency care," Eur. Heart J. Digit. Health, vol. 5, pp. 454–460, 2024.
[2]. A. H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," Nat. Commun., vol. 11, no. 1, p. 1760, Apr. 2020.
[3]. A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nat. Med., vol. 25, no. 1, pp. 65–69, Jan. 2019.
[4]. N. Strodthoff, P. Wagner, M. Schaeffter, and W. Samek, "Deep learning for ECG analysis: A review," IEEE J. Biomed. Health Inform., vol. 25, no. 5, pp. 1512–1525, May 2021.
[5]. M. A. Reyna et al., "Will two do? Varying dimensions in electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021," Comput. Cardiol., vol. 48, pp. 1–4, Sept. 2021.
[6]. Y. Zhou et al., "Masked transformer for electrocardiogram classification," arXiv preprint arXiv:2309.07136, 2023.

[7]. E. Perez-Alday et al., "The PhysioNet/Computing in Cardiology Challenge 2021 dataset," PhysioNet, 2021.
[8]. A. Ukil, G. Kulkarni, A. Ghose, and R. S. Mahapatra, "Resource-constrained CVD classification using single-lead ECG on wearable and implantable devices," IEEE J. Biomed. Health Inform., vol. 26, no. 4, pp. 1418–1428, 2022.
[9]. C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," arXiv preprint arXiv:1706.02633, 2017.
[10]. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[11]. J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," Int. J. Comput. Vis., vol. 129, no. 6, pp. 1789–1819, 2021.
[12]. X. An, S. Shi, Q. Wang, Y. Yu, and Q. Liu, "Research on a lightweight arrhythmia classification model based on knowledge distillation for wearable single-lead ECG monitoring systems," Sensors, vol. 24, no. 24, Art. no. 7896, 2024.
[13]. A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Enhancing Breast Cancer Screening in Kenya Through Deep Learning-Driven Mammographic Analysis

Edna C. Too
Computer Science Department
Chuka University
Chuka, Kenya
echebet@chuka.ac.ke
ORCID:0000-0001-5885-4840

David G Mwathi
Computer Science Department
Chuka University
Chuka, Kenya
dgmwathi@chuka.ac.ke
ORCID:0009-0001-6405-9170

Pauline W. Mwaka
Computer Science Department
Chuka University
Chuka, Kenya
pmwaka@chuka.ac.ke
ORCID:0009-0008-5824-5170

Saif Kinyori
Computer Science Department
Chuka University
Chuka, Kenya
skinyori@chuka.ac.ke
ORCID:0009-0005-8833-0091

Julius Gachoki
Medical Department
Chuka University
Chuka, Kenya
jgachoki@chuka.ac.ke

Mary Wanjiku
Computer Science Department
Chuka University
Chuka, Kenya
mshiku555@gmail.com

Lucy Gitonga
Nursing Department
Chuka University
Chuka, Kenya
lgitonga@chuka.ac.ke
0009-0006-8937-013X

Abstract
Breast cancer is one of the most prevalent cancers globally among women and remains a leading cause of death caused by cancer. A major way to improve patient survival rates is to detect cancer early, before it spreads to other parts of the body. However, mammographic screening relies heavily on radiologists, thereby making the diagnostic process time-consuming and susceptible to interpretive variability. In Kenya, these challenges are further aggravated by a shortage of radiologists, often resulting in delayed diagnosis and treatment. This study developed an automated breast cancer screening system using mammographic X-ray images and deep learning techniques. An experimental quantitative research design was adopted to evaluate three families of deep learning architectures: Capsule Neural Networks, Vision Transformer architectures, and Convolutional Neural Networks. The models were trained using 90% of the VinDr-Mammo dataset and evaluated on a locally collected dataset obtained from level 5 and level 6 hospitals in Kenya. Performance was assessed using accuracy, precision, recall, and F1 score metrics. According to the results, CapsNet achieved the highest accuracy of 90.09% with a recall of 89%. The vision transformer model achieved comparable recall and F1score performance while maintaining a significantly smaller model size of 12.3 MB. The findings suggest that both CapsNet and Vision Transformer models have significant potential for supporting breast cancer screening in resource-constrained healthcare environments. Their adoption could improve access to quality diagnostic services, facilitate earlier detection, reduce diagnostic delays, and ultimately improve patient outcomes in Kenya.

Keywords: Breast cancer; Mammography; Automated detection; Deep learning; Convolutional Neural Networks; Capsule Networks



I. INTRODUCTION

Breast cancer is one of the leading global health concerns and continues to be the most frequently diagnosed cancer in women [1]. Despite advances in prevention, it accounts for a substantial portion of cancer-related deaths. In 2022, breast cancer accounted for an estimated 2.3 million new cases worldwide and approximately 670,000 deaths [2]. If the current trend continues, the global burden of the disease is expected to rise substantially, with new cases projected to increase by 38% and breast cancer-related deaths by 68% by 2050 [3]. Early detection is essential for improving breast cancer survival and treatment outcomes. Despite its importance, access to screening services varies considerably across countries, with low and middle-income countries facing greater challenges than high-income countries, therefore leading to disparities in health outcomes [4]. In Kenya, breast cancer remains a significant public health challenge and contributes substantially to the country's disease burden. The Ministry of Health, Kenya, reported an age-standardized incidence rate (ASR) of 41 cases and a mortality rate of 19.4 deaths per 100,000 women [5]. Despite ongoing efforts to control the disease, breast cancer is responsible for approximately 3,100 deaths annually [6]. Several interconnected factors continue to hinder effective breast cancer detection in Kenya. Access to screening services remains limited, particularly in rural regions where mammography facilities and specialized diagnostic services are scarce [7]. Furthermore, the cost of screening, diagnosis, and treatment poses a significant barrier, often resulting in delayed healthcare-seeking [8]. Traditional breast cancer screening, primarily through mammography, also relies heavily on manual interpretation by radiologists, making the diagnostic process time-consuming and dependent on expertise [9]. In Kenya, the shortage of trained radiologists further aggravates this challenge, contributing to delayed diagnosis and, consequently, poor patient outcomes [10]. Advances in deep learning offer a promising opportunity to improve breast cancer detection by enabling automated feature extraction and accurate classification of medical images [11]. Such approaches have the potential to support radiologists by reducing diagnostic workload and facilitating early disease detection, thereby improving survival rates. This study, therefore, focused on developing an automated breast cancer detection system using mammographic X-ray images. Specifically, it conducted a comparative evaluation of Capsule Networks (CapsNets), Vision Transformers (ViTs), and Convolutional neural networks (CNNs) for the detection and classification of breast cancer tissues. The aim was to identify the most effective model for accurate and reliable breast cancer diagnosis. The findings are expected to provide evidence for the adoption of deep learning-based diagnostic tools for breast cancer screening and diagnosis in Kenyan hospitals.

II. RELATED WORK

In recent years, deep learning has been widely used to address challenges such as tuberculosis detection [12], lung cancer detection [13], masked face recognition [14], and early breast cancer detection [15]. Among the various deep learning approaches, Convolutional Neural Networks (CNNs) have emerged as one of the most widely adopted architectures for medical image analysis due to their ability to learn hierarchical spatial features from images [16] automatically. For example, a CNN-based cancer detection system developed in [17] achieved accuracies of 86.42% on the BreakHis dataset and 98.97% on the MIAS dataset. Similarly, Yuan et al. [18] conducted a comparative evaluation of fourteen deep learning models on the BreakHis dataset and reported excellent performance, with ConvNeXT achieving 99.2% accuracy in binary classification. However, the accuracy dropped to 94.2% with multiclass classification. Lightweight CNN models such as LBNet have also demonstrated strong performance while reducing computational complexity, achieving 97.28% accuracy [19]. Despite their success, CNNs generally require large training datasets and may struggle to capture long-range contextual relationships within images. To overcome some of these limitations, researchers have explored Capsule Networks (CapsNets), which preserve spatial relationships between image features and can perform effectively with limited training data. For instance, UCapsNet combined segmentation and classification to

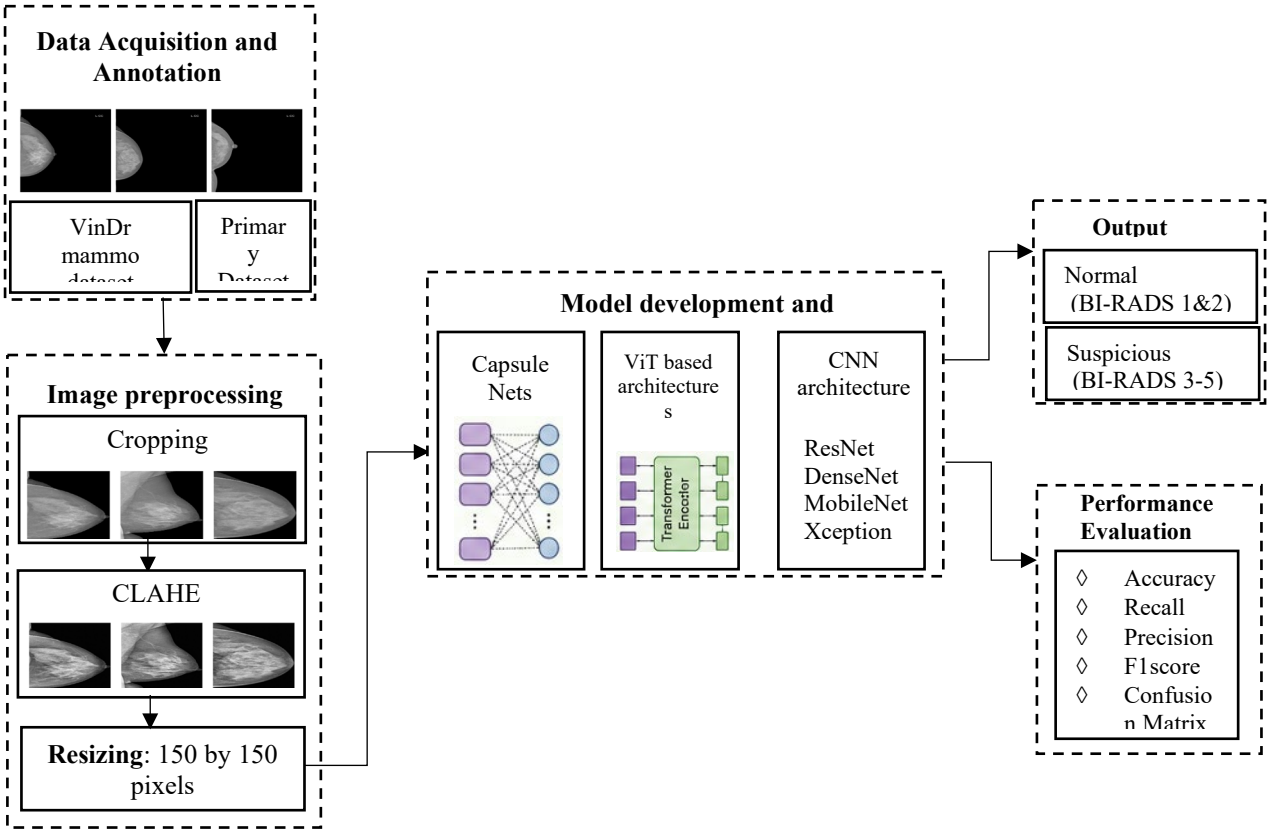


achieve 99.22% accuracy in breast ultrasound image analysis [15]. Similarly, TTDCapsNet reported an accuracy of approximately 95.01% on breast cancer datasets, demonstrating competitive performance against several state-of-the-art approaches [20]. However, CapsNets often require architectural enhancements to handle complex medical images effectively. More recently, transformers have gained attention in breast cancer detection for their ability to capture global relationships in images using self-attention mechanisms. A recent study combined CNNs and Vision Transformers for mammogram classification, achieving 90.1% accuracy on CLAHE-enhanced mammographic images [21]. Similarly, a study evaluating transformer-based models, including Swin Tiny, DeiT, BEiT, and ViT, reported that the ViT model achieved the highest accuracy of 99.32% on mammographic image datasets [22]. Despite these promising results, transformer-based models typically require substantial computational resources and large training datasets. Although CNNs, CapsNets, and Transformers have all demonstrated promising results in breast cancer detection, most studies evaluate these architectures independently, making direct comparison difficult. Furthermore, the majority of existing studies rely on publicly available datasets that largely originate from non-African populations. Consequently, there is limited evidence regarding the relative performance of these architectures and their suitability for deployment in Kenyan healthcare settings. This study addresses this gap by conducting a unified comparative evaluation of CNNs, Capsule Networks, and Vision Transformers for breast cancer detection, aiming to identify the most effective architecture to support breast cancer diagnosis in Kenyan hospitals.

III. METHODOLOGY

1. Research Design

The study adopted an experimental quantitative research design to develop and evaluate deep learning models for automated breast cancer detection using mammographic X-ray images. It conducted a comparative assessment of Capsule Networks (CapsNets) and vision transformer architectures, and Convolutional Neural Networks (CNNs), including ResNet, DenseNet, MobileNet, EfficientNet, and Xception.



1. Figure 1.Research Design

The objective was to identify the most accurate and efficient architecture for breast cancer detection and provide evidence on its suitability for deployment in Kenyan healthcare settings. The study was conducted in four stages: data acquisition and annotation, image preprocessing, model development and training, and performance evaluation. Mammographic images were obtained from both the publicly available VinDr-Mammo dataset and selected level 5 and 6 hospitals in Kenya. The images were preprocessed to enhance quality and ensure consistency before being used to train CapsNet, Transformer, and CNN-based models for breast cancer classification. The performance of the models was evaluated using accuracy, precision, recall, and F1 Score to identify the most effective. architecture for breast cancer detection. Figure 1 shows the research design.

2. X-ray Image Acquisition

The study used two mammographic datasets for model development and evaluation. The first dataset was the publicly available VinDr-Mammo dataset, which consists of 20,000 full-field digital mammograms (FFDM) annotated by expert radiologists and labeled according to the Breast Imaging Reporting and Data System (BI-RADS) scores ranging from 1 to 5, with higher categories indicating a greater likelihood of malignancy. Figure 2 shows the distribution of images per BI-RAD class. The second dataset comprised 8,789 locally collected and curated mammographic images from selected level 5 and level 6 hospitals, namely Kerugoya County Referral Hospital, Nyeri County Referral Hospital, Meru Level 5 Hospital, Outspan Hospital, and Embu Level 5 Hospital. The images were anonymized and annotated by certified radiologists in accordance with BI-RADS standards. Figure 3 shows the distribution of the dataset per BIRAD classes. The local dataset was used to validate the models and improve their applicability in Kenyan healthcare settings. Due to the substantial class

imbalance across the BI-RADS categories, as shown in Figures 2 and 3, BI-RADS 3, 4, and 5 were combined into a single suspicious class, while BI-RADS 1 and 2 were grouped as normal/benign. This binary classification was adopted to provide more balanced sample sizes and support stable model training and evaluation compared to a multiclass classification approach. Similar grouping strategies have been applied in previous breast imaging machine learning studies [23]. The classification was intended to distinguish normal or benign examinations from those requiring further assessment or diagnostic work-up. However, BI-RADS 3, 4, and 5 have different malignancy risks and clinical management recommendations. Therefore, the combined class represents a machine learning classification decision and should not be interpreted as indicating the same level of clinical risk across these categories.

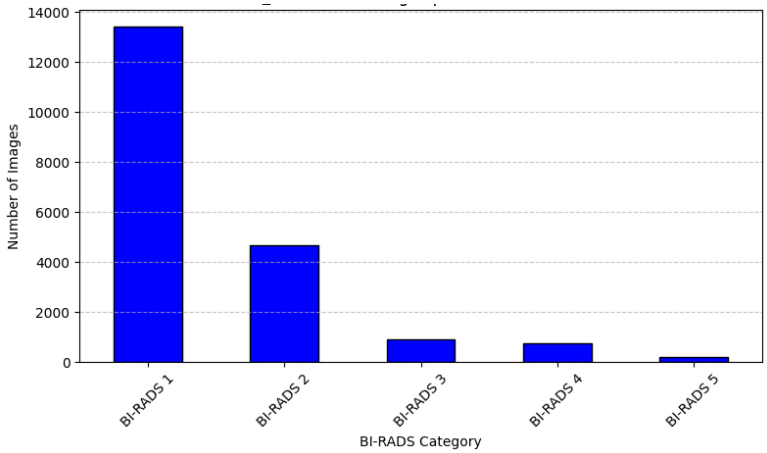


Figure 2. Distribution of Vin Dr-Mammo Dataset per BI-RAD Class

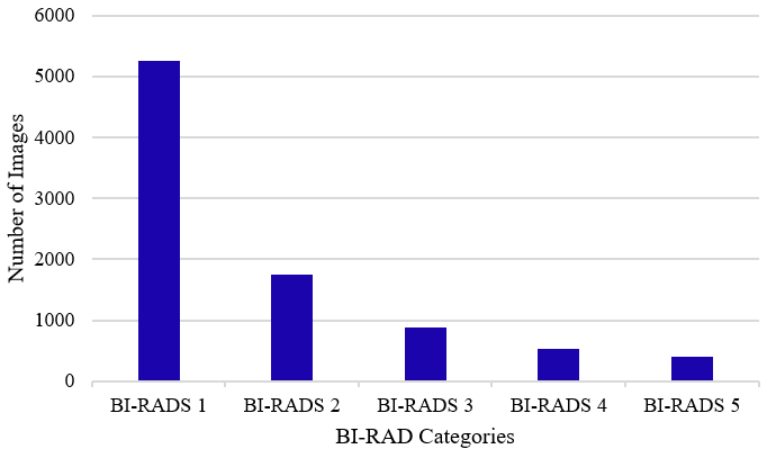
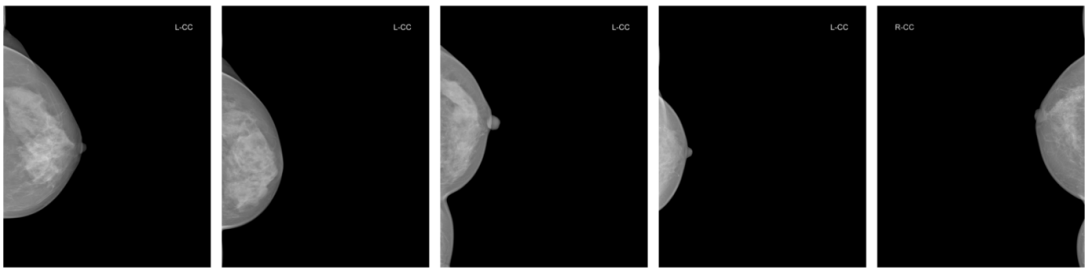


Figure 3. Distribution of the locally collected Kenyan Dataset per BI-RAD Class

To ensure image quality and uniformity, preprocessing involved cropping each mammogram to isolate the breast region, as shown in Figure 4. Afterward, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to the cropped mammograms to enhance local contrast, as shown in Figure 5. Finally, all the

preprocessed images were resized to 150 × 150 pixels to standardize input dimensions and reduce computational complexity. No additional preprocessing techniques were applied to ensure that the models learn from realistic



clinical data. The VinDr-Mammo dataset was then divided into training (90%) and validation (10%) subsets for model development, while the locally collected mammographs from Kenyan hospitals were reserved exclusively for testing.

Figure 4. Sample Breast Mammogram Images before preprocessing

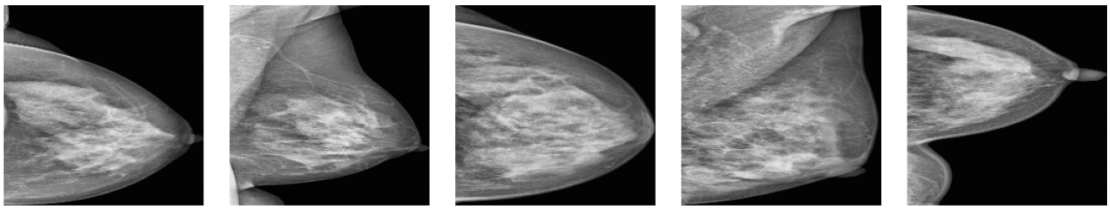


Figure 5. Breast Mammogram images after cropping and applying CLAHE

3. Feature Extraction and Classification

The study evaluated three categories of deep learning architectures for breast cancer detection: Capsule Neural Networks (CapsNets), vision transformers, and Convolutional Neural Networks (CNNs). CapsNets, introduced by Hinton et al. [24], were selected for their ability to preserve spatial hierarchies and positional relationships among image features. Unlike conventional CNNs, CapsNets represent features as multidimensional vectors and employ a dynamic routing mechanism that maintains meaningful relationships between lower and higher-level features. The architecture also incorporates a reconstruction mechanism that improves regularization and model interpretability. Vision Transformers (ViTs) were evaluated for their ability to model global contextual relationships within mammographic images through self-attention mechanisms. By capturing long-range dependencies between image regions, ViT models can effectively identify distributed abnormalities that may not be easily detected through local feature extraction alone [21]. To provide benchmark performance, five widely used CNN architectures were implemented: ResNet, DenseNet, MobileNet, EfficientNet, and Xception. ResNet and DenseNet were selected for their deep feature extraction capabilities and efficient gradient propagation, while MobileNet and EfficientNet were chosen for their lightweight, computationally efficient designs. Xception was included due to its use of depth-wise separable convolutions, which improve computational efficiency while maintaining high classification performance. The same preprocessing procedures, training parameters, and evaluation metrics were applied to all models to enable a consistent and unbiased comparison of their performance.

4. Model Training and Evaluation

Model training and evaluation took place in a Linux environment with an NVIDIA A30 GPU, 24GB of RAM, 64GB of RAM, 1TB SSD, and 24 vCPUs. All models were trained using the Adam optimizer, a batch size of 32, 100 epochs, and a learning rate of 1e-4. To minimize overfitting and enhance model generalization, early stopping was incorporated throughout training. Model performance was validated using the local dataset collected from Level 5 and Level 6 hospitals in Kenya. The evaluation assessed the models' ability to classify mammograms according to BI-RADS categories accurately and to distinguish between normal and abnormal breast tissues. A confusion matrix was used to compare predicted and actual labels, summarizing results as true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). These values were subsequently used to compute performance metrics: accuracy, precision, recall, and F1 score. Accuracy was used to determine the overall proportion of correctly classified mammograms. Precision measured the proportion of positive predictions that were correctly identified, while recall evaluated the models' ability to detect positive breast cancer cases. The F1 score, which represents the harmonic mean of precision and recall, provided a balanced assessment of model performance predictions, particularly in the presence of class imbalance.

IV. RESULTS AND DISCUSSION

1. Training and Validation Performance

The trained models were evaluated using training and validation accuracy and loss over 100 epochs. MobileNet demonstrated rapid convergence with high accuracy and minimal overfitting, as shown in Figure 10. At the same time, DenseNet and Xception exhibited stable learning behavior throughout training, achieving consistent performance across epochs, as illustrated in Figure 11 and Figure 8, respectively. ResNet achieved competitive accuracy but exhibited fluctuations in validation performance, suggesting sensitivity to hyperparameter settings, as shown in Figure 9. CapsNet demonstrated strong learning and generalization capabilities, characterized by high validation and a small gap between training and validation performance, as shown in Figure 6. Similarly, the vision transformer-based model maintained stable validation metrics throughout training, suggesting good generalization to unseen data, as illustrated in Figure 7.

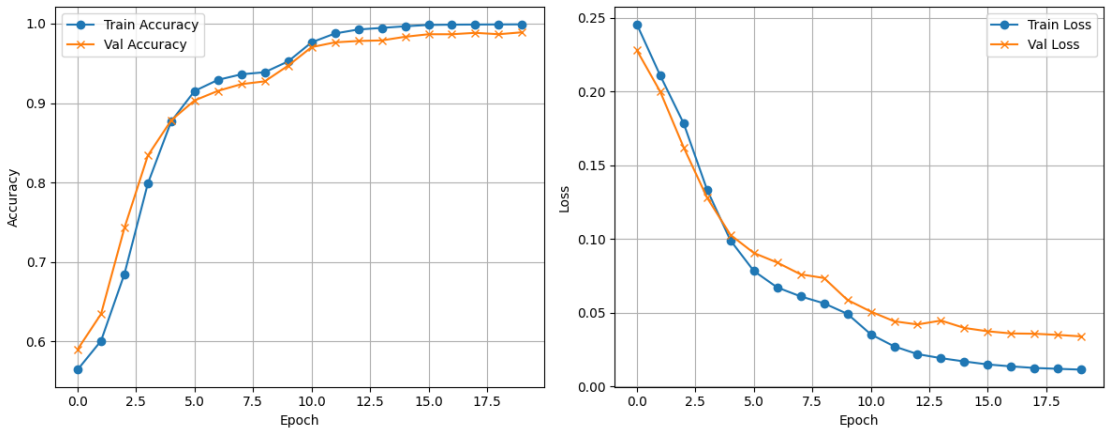


Figure 6. Capsule Nets training and validation accuracy and loss curves

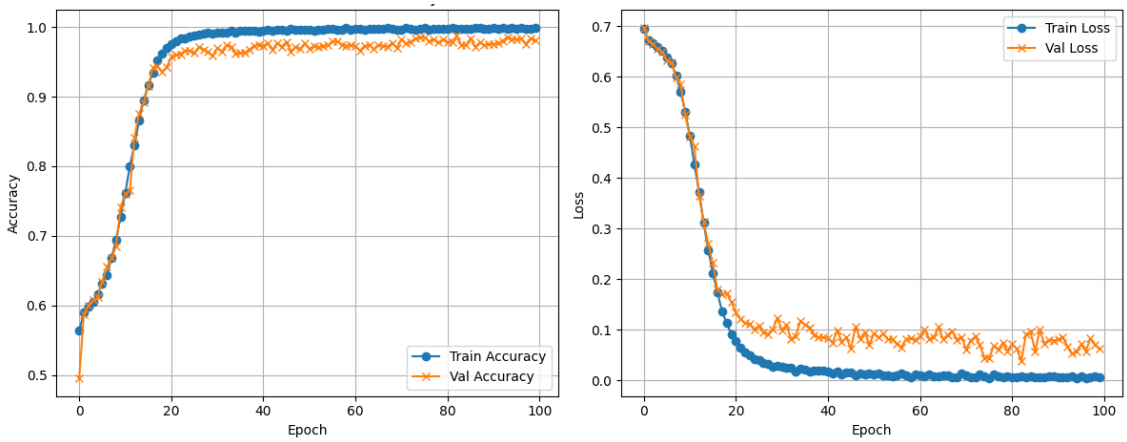


Figure 7. Vision Transformer training and validation accuracy and loss curves

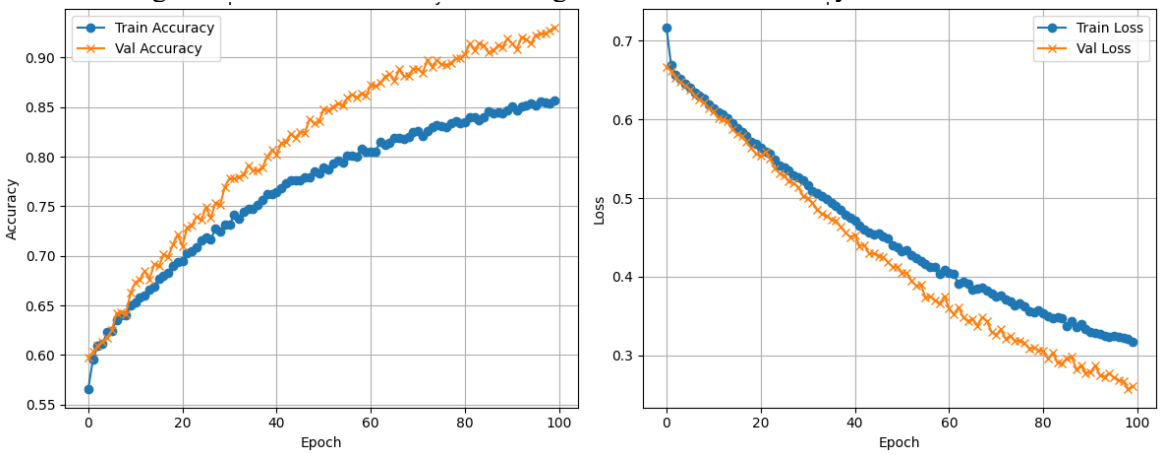


Figure 8. Xception training and validation accuracy and loss curves

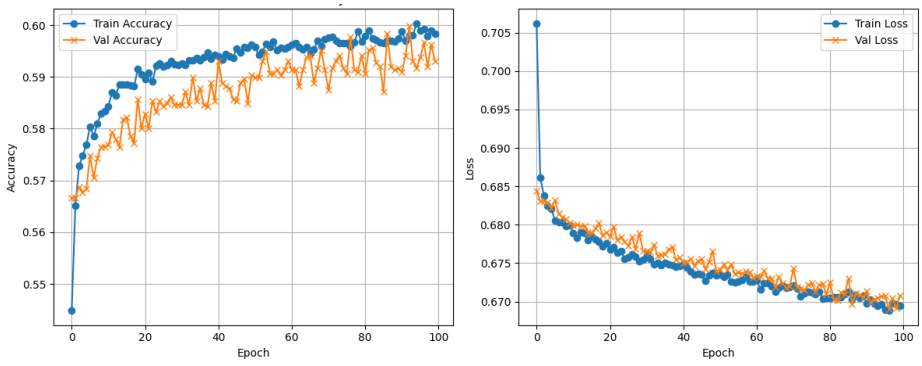


Figure 9. ResNet training and validation accuracy and loss curves

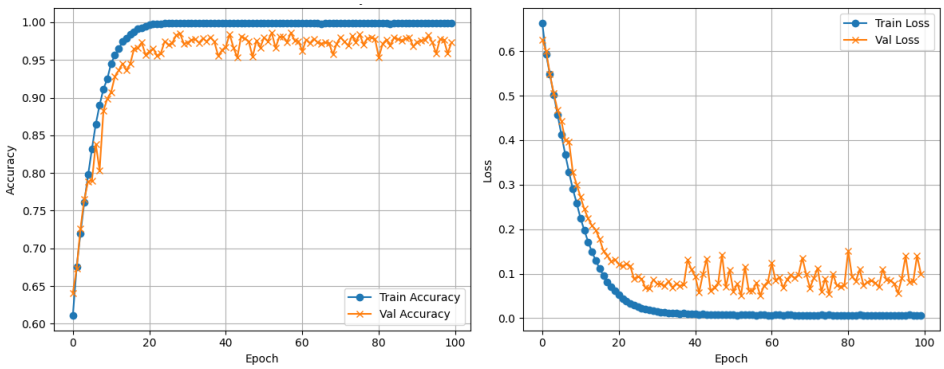


Figure 10. MobileNet training and validation accuracy and loss curves

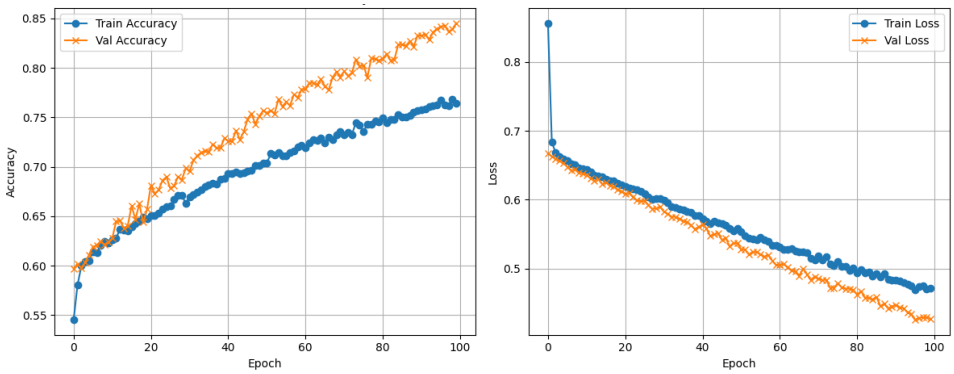


Figure 11. DenseNet training and validation accuracy and loss curves

2. Confusion Matrix

A confusion matrix was generated for each model to assess classification performance across the BI-RADS categories. CapsNet achieved the highest number of correctly classified benign cases (1771) and the lowest number of false positives (37), indicating strong specificity, as shown in Figure 12. ResNet and the vision transformer model showed similar performance, correctly classifying most benign cases but detecting fewer suspicious cases, as shown in Figure 13 and Figure 10, respectively. MobileNet achieved a more balanced performance by identifying more suspicious cases while maintaining a relatively low false-positive rate, as shown in Figure 11. In contrast, DenseNet and Xception recorded higher misclassification rates, as shown in Figures 14 and 15. Overall, CapsNet excelled in reducing false alarms, while MobileNet achieved a better balance between sensitivity and specificity.

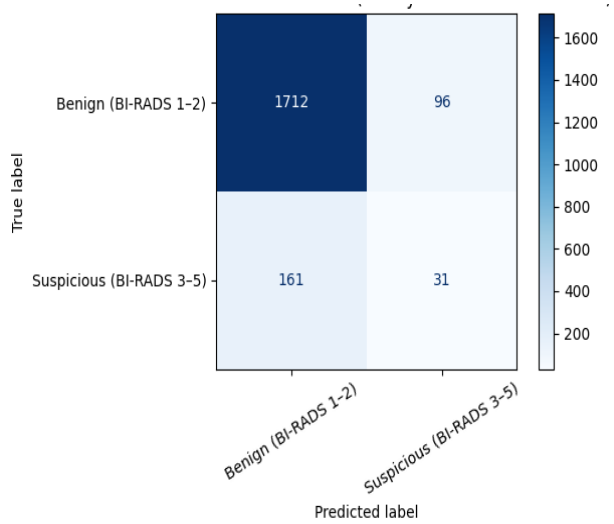


Figure 12. MobileNet Confusion Matrix

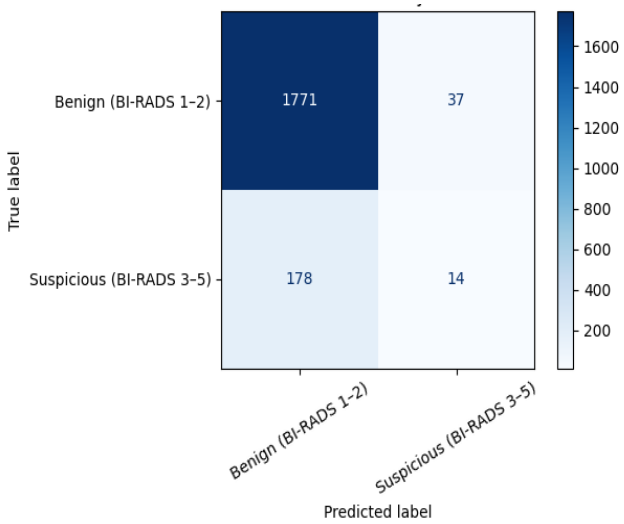


Figure 13. CapsNet Confusion Matrix

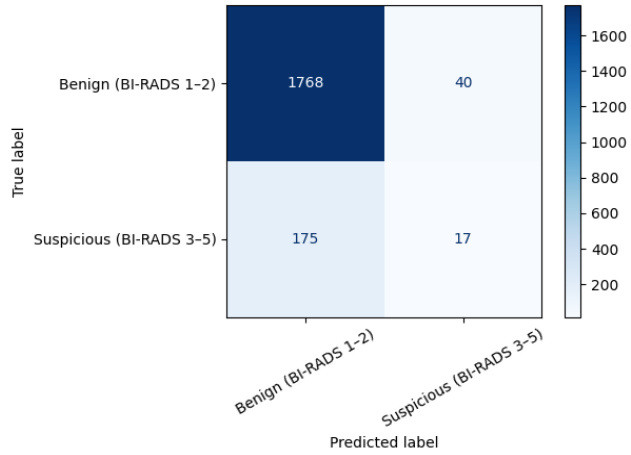


Figure 14. Vision Transformer Confusion Matrix

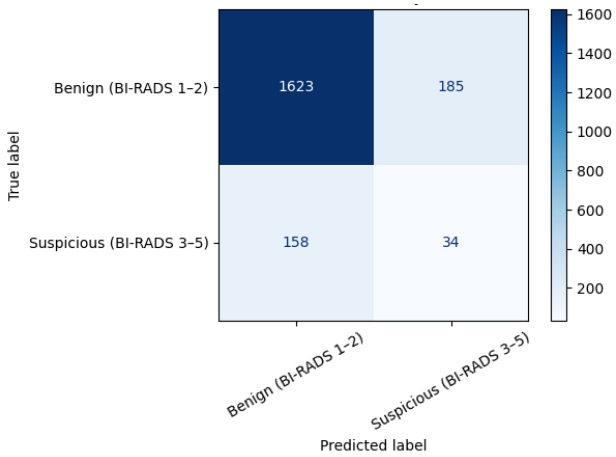


Figure 15. DenseNet Confusion Matrix

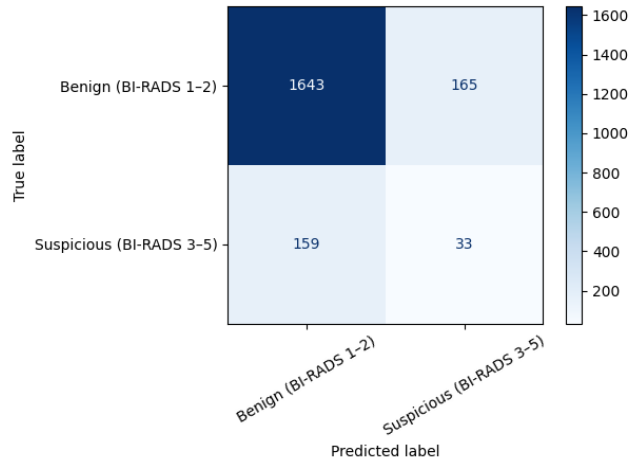


Figure 16. Xception Confusion Matrix

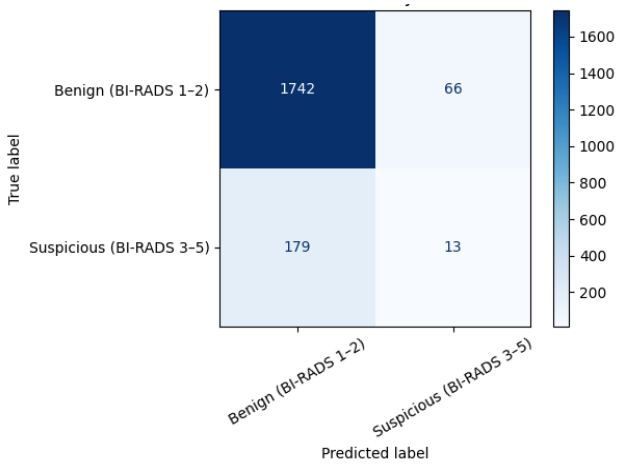


Figure 17. ResNet Confusion Matrix

3. Model Performance Evaluation

The evaluation results indicated that CapsNet achieved the highest overall accuracy of 90.09%, accompanied by a strong recall score. These results suggest that CapsNet is effective at identifying malignant cases and capturing relevant spatial relationships in mammographic images. The Vision Transformer model followed closely, with comparable recall and F1-score performance, while maintaining a smaller model size of 12.3 MB. Therefore, demonstrating its potential as a computationally efficient alternative for breast cancer detection, as shown in Table 1. Baseline CNN and ResNet models demonstrated moderate but consistent performance, although they required larger model sizes. MobileNet provided a favorable balance between classification accuracy and computational efficiency, making it suitable for deployment in resource-constrained environments. In contrast, DenseNet and Xception recorded lower performance, which may be attributed to limited generalization capability or overfitting to the training data. Overall, the findings suggest that CapsNet and Vision Transformer architectures offer improved classification performance and model efficiency compared to conventional CNN-based approaches for mammographic breast cancer detection.

Table 1. A Performance Comparative Table for all Trained Models

Model	Params (M)	Size (MB)	Acc (%)	Loss	Prec (%)	Rec (%)	F1 (%)
CapsNet	5.29	60.6	90.09	0.10	85	89	86
ViT	0.07	12.3	89.90	1.15	85	89	87
CNN	11.17	55.3	89.24	1.47	84	89	86
ResNet	0.13	93.6	88.43	0.55	84	88	85
MobileNet	3.23	14.4	87.76	0.90	85	87	86
Xception	0.13	83.0	84.47	0.43	84	84	84
DenseNet	0.13	29.6	83.84	0.46	84	83	83

4. Model Prototype

A prototype of the proposed breast cancer screening system was developed and deployed on the Hugging Face Platform. The prototype was subsequently evaluated using mammographic images collected from selected Kenyan hospitals. It was designed to classify BI-RADS categories 1 and 2 as normal and BI-RADS categories 3 to 5 as suspicious. Figures 18 and 19 illustrate the operation of the prototype for both classification outcomes.

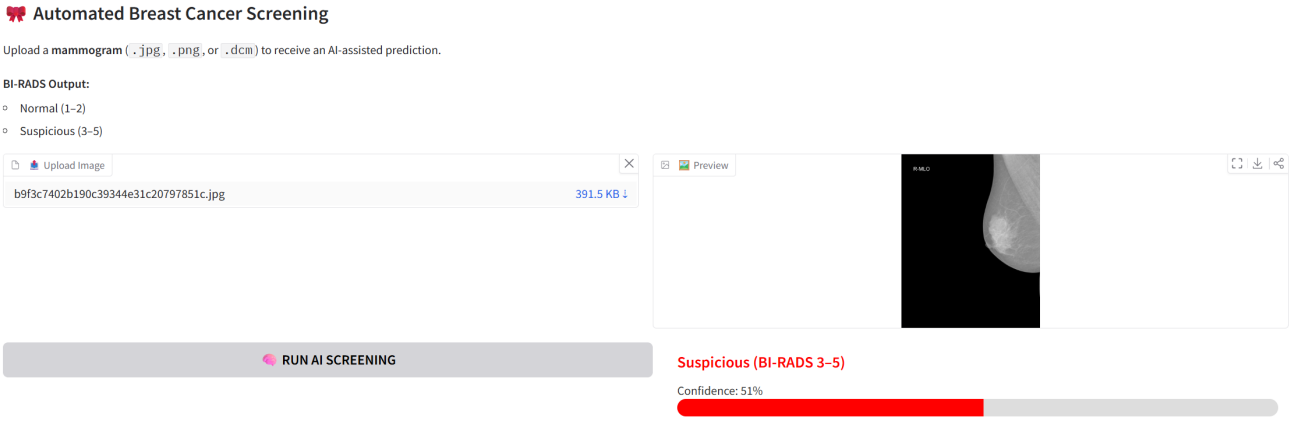


Figure 18. Prototype performance on a normal mammographic image

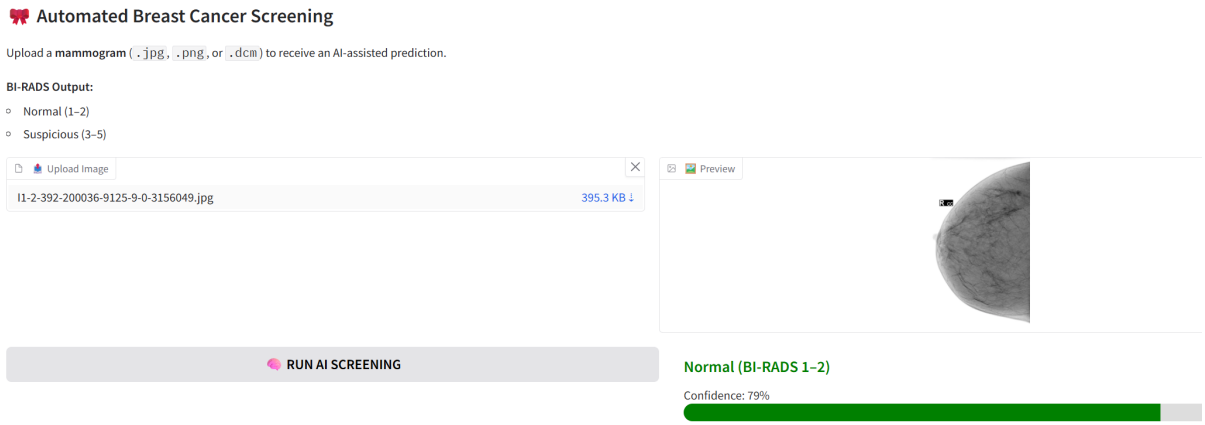


Figure 19. Prototype performance on a suspicious mammographic image

V. CONCLUSION

This study compared Convolutional Neural Networks (CNNs), Vision Transformers, and Capsule Networks for breast cancer detection using mammographic images. The results showed that CapsNet achieved the highest overall accuracy of 90.09%, demonstrating strong capability in capturing spatial relationships and identifying malignant cases. The Vision Transformer model achieved comparable performance while maintaining a smaller model size. Among the CNN architectures, MobileNet provided the best balance between classification performance and computational efficiency. Overall, the findings suggest that CapsNet and Vision Transformer-based architectures offer promising approaches for breast cancer detection, while lightweight CNNs remain practical alternatives where computational resources are limited.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the National Cancer Institute of Kenya (NCI-Kenya) for funding this research. The authors also thank the Kenya Education Network (KENET) for providing the computational infrastructure and resources that supported this study.

ETHICAL CONSIDERATIONS

Ethical approval for the use of locally collected mammographic data was obtained from the relevant participating hospitals and the Kenya Ministry of Health. All patient data were anonymized before analysis and handled in accordance with medical privacy regulations. The publicly available VinDr-Mammo dataset was used in compliance with its open-access research guidelines. Throughout the study, principles of research integrity, confidentiality, data security, and responsible use of Artificial intelligence in healthcare were strictly observed.

REFERENCES

[1]. J. Kim et al., "Global patterns and trends in breast cancer incidence and mortality across 185 countries," Nat Med, vol. 31, no. 4, pp. 1154–1162, Apr. 2025, doi: 10.1038/s41591-025-03502-3.
[2]. World Health Organization, "Breast cancer." Accessed: Jun. 22, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
[3]. S. Nature, "Estimates of Global Breast Cancer Mortality and Incidence Are on the Rise," Today's Clinical Lab. Accessed: Jun. 22, 2026. [Online]. Available: https://www.clinicallab.com/estimates-of-global-breast-cancer-mortality-and-incidence-are-on-the-rise-28185
[4]. C. H. Barrios, "Global challenges in breast cancer detection and treatment," The Breast, vol. 62, pp. S3–S6, Mar. 2022, doi: 10.1016/j.breast.2022.02.003.
[5]. Ministry of Health, "NATIONAL_CANCER_MEAL_FRAMEWORK_2023-2027," Ministry of Health, 2023. Accessed: Jun. 19, 2026. [Online]. Available: https://www.iccp-portal.org/sites/default/files/resources/NATIONAL_CANCER_MEAL_FRAMEWORK_2023-2027.pdf
[6]. B. Hutchinson et al., "An economic evaluation of breast cancer interventions in Kenya," eClinicalMedicine, vol. 77, p. 102894, Nov. 2024, doi: 10.1016/j.eclinm.2024.102894.
[7]. F. Wambalaba and A. Wambalaba, "The Inequity of Cancer Treatment Services in Kenya," in The Landscape of Global Health Inequity, vol. 22, B. W. K. Son, Ed., in Integrated Science, vol. 22. , Cham: Springer Nature Switzerland, 2024, pp. 37–46. doi: 10.1007/978-3-031-60502-4_4.
[8]. R. Gakunga et al., "Identifying Barriers and Facilitators to Breast Cancer Early Detection and Subsequent Treatment Engagement in Kenya: A Qualitative Approach," The Oncologist, vol. 24, no. 12, pp. 1549–1556, Dec. 2019, doi: 10.1634/theoncologist. 2019-0257.
[9]. N. K. Kimario, "Prediction of Breast Cancer Using Mammogram Images through Deep Learning Techniques," East Afr. j. eng., vol. 8, no. 2, pp. 453–467, Nov. 2025, doi: 10.37284/eaje. 8.2.4040.
[10]. D. J. Mollura et al., "Artificial intelligence in low-and middle-income countries: innovating global health radiology," Radiology, vol. 297, no. 3, pp. 513–520, 2020.
[11]. A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: a comprehensive review," Multimedia tools and applications, vol. 84, no. 21, pp. 24079–24190, 2025.
[12]. E. C. Too, D. G. Mwathi, L. K. Gitonga, S. Kinyori, and P. Mwaka, "A Pruned and Quantized Convolution Neural Network for Enhancing Tuberculosis Detection Using Chest X-Ray Images," in

Measurement, Automation, and Control in Artificial Intelligence and Machine Learning, Springer, 2025, pp. 33–68.
[13]. T. Meeradevi, S. Sasikala, L. Murali, N. Manikandan, and K. Ramaswamy, "Lung cancer detection with machine learning classifiers with multi-attribute decision-making system and deep learning model," Sci Rep, vol. 15, no. 1, p. 8565, Mar. 2025, doi: 10.1038/s41598-025-88188-w.
[14]. P. W. Mwaka, D. G. Mwathi, and E. C. Too, "Fine-Tuned Convolutional Neural Network and Linear Multiclass Support Vector Machine for Hybrid Masked Face Recognition," in Measurement, Automation, and Control in Artificial Intelligence and Machine Learning, Springer, 2025, pp. 69–83.
[15]. G. Madhu et al., "UCapsNet: A Two-Stage deep learning model using U-Net and capsule network for breast cancer segmentation and classification in ultrasound imaging," Cancers, vol. 16, no 22, p. 3777, 2024.
[16]. F. Nasir, S. Rahman, and N. Nasir, "Breast Cancer Detection Using Convolutional Neural Networks: A Deep Learning-Based Approach," Cureus, May 2025, doi: 10.7759/cureus. 83421.
[17]. K. Aguerchi, Y. Jabrane, M. Habba, M. Ameur, and A. H. E. Hassani, "Enhancing Automated Breast Cancer Detection: A CNN-Driven Method for Multi-Modal Imaging Techniques," JPM, vol. 15, no. 10, p. 467, Oct. 2025, doi: 10.3390/jpm15100467.
[18]. B. Yuan et al., "Comparative analysis of convolutional neural networks and transformer architectures for breast cancer histopathological image classification," Frontiers in Medicine, vol. 12, p. 1606336, 2025.
[19]. J. Ahmmed et al., "LBNet: an optimized lightweight CNN for mammographic breast cancer classification with XAI-based interpretability," Scientific Reports, 2025.
[20]. V. Akoto-Adjepong, O. Appiah, P. K. Mensah, and P. Appiahene, "TTDCapsNet: Tri Texton-Dense Capsule Network for complex and medical image recognition," Plos one, vol. 19, no. 3, p. e0300133, 2024.
[21]. V. R. Patheda, G. Laxmisai, B. V. Gokulnath, S. P. Siddique Ibrahim, and S. Selva Kumar, "A Robust Hybrid CNN+ViT Framework for Breast Cancer Classification Using Mammogram Images," IEEE Access, vol. 13, pp. 77187–77195, 2025, doi: 10.1109/ACCESS.2025.3563218.
[22]. S. Osman et al., "Fine-tuned Transformer Models for Breast Cancer Detection and Classification," 2025, arXiv. doi: 10.48550/ARXIV.2512.02091.
[23]. A. Sabani et al., "BI-RADS-based classification of mammographic soft tissue opacities using a deep convolutional neural network," Diagnostics, vol. 12, no. 7, p. 1564, 2022.
[24]. G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in International Conference on Learning Representations, 2018.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Territorial Crime Pattern Classification in Peru Using Unsupervised Learning and K-Medoids, 2015–2024

Renzo Alberto Taco Coayla
Universidad Privada de Tacna
Tacna, Perú
rentaco@upt.pe
https://orcid.org/0000-0003-0014-7216

Martha Judith Paredes Vignola
Universidad Privada de Tacna
Tacna, Perú
majparedes@upt.pe
https://orcid.org/0000-0002-5046-0908

Enrique Félix Lanchipa Valencia
Universidad Privada de Tacna
Tacna, Perú
elanchipa@upt.pe
https://orcid.org/0000-0001-8164-9781

Elard Ricardo Rodríguez Marca
Universidad Privada de Tacna
Tacna, Perú
elrrodriguez@upt.pe
https://orcid.org/0000-0002-1220-1972

Hugo Martín Alcántara Martínez
Universidad Privada de Tacna
Tacna, Perú
halcantaram@upt.pe
https://orcid.org/0000-0003-3223-7929

Josué Abraham Emanuel Chambilla Zúñiga
Universidad Privada de Tacna
Tacna, Perú
jaechambilla@upt.pe
https://orcid.org/0009-0000-5080-841X

Abstract
Registered crime exhibits a heterogeneous and multidimensional territorial distribution that cannot be adequately explained through absolute counts or administrative rankings alone. This study aims to classify and characterize territorial crime patterns in Peru using unsupervised learning and K-Medoids for the period 2015–2024. A balanced analytical matrix was built with 250 region-year observations, covering 25 regions and 12 crime categories expressed as rates per 100,000 inhabitants. Data were obtained from DATACRIM–INEI, harmonized by territory, converted into population-adjusted rates, and transformed into a multivariate machine-learning matrix. K-Medoids was applied using Manhattan distance and robust scaling. Cluster solutions from k = 2 to k = 10 were evaluated through the silhouette coefficient, Davies-Bouldin index, Calinski-Harabasz index, and Gap Statistic. The variation of internal validation metrics identified k = 10 as the optimal clustering solution, achieving the best overall performance across the Davies–Bouldin index, Calinski–Harabasz index, and Gap Statistic. Longitudinal validation for the 2015–2021 period yielded an Adjusted Rand Index (ARI) of 0.6046, confirming a statistically significant and stable structural agreement over time. The final partition revealed ten distinct granular region-year crime profiles, isolating specific dynamics associated with public safety, property-patrimonial offenses, domestic conflict, gender-based violence, and drug markets. This study contributes a replicable framework for classifying territorial crime in Latin American contexts through official data, normalized crime rates, and interpretable unsupervised learning.

Keywords: crime rates; DATACRIM; K-Medoids; Peru; public safety; territorial crime; unsupervised learning

I. INTRODUCTION

Crime is a territorially heterogeneous, dynamic, and multidimensional phenomenon whose distribution cannot be understood solely through absolute counts or administrative rankings [1]. Criminological evidence has shown that crimes tend to be spatially concentrated in certain territories and that this concentration responds to conditions of opportunity, mobility, exposure to risk, availability of victims, and the presence or absence of

formal and informal control mechanisms. In this sense, the analysis of crime requires approaches capable of identifying distinct patterns across territorial units [2], especially in countries with high geographic, social, and institutional diversity, such as Peru. The inadequacy of traditional classifications in the Peruvian and regional contexts lies in their inability to address the complexity of criminal governance, a phenomenon in which armed actors establish mechanisms of social and economic regulation that coexist with the State. This dynamic generates territorial criminal "signatures" where daily life is mediated by illegal actors in vulnerable areas [3]. This demonstrates that urban victimization in Peru is not uniform but follows distinct spatial patterns determined by the size of population centers. In large cities, high density increases opportunities for crime. At the same time, in medium and small centers, community cohesion can act as a protective factor, even generating a negative correlation between population growth and crime [4]. Recent developments in machine learning have expanded the analytical possibilities in quantitative criminology. Models such as Random Forest, Support Vector Machines, neural networks, LSTM, ConvLSTM, and deep spatiotemporal architectures have been used to predict risk zones, map hotspots, and model rare or unbalanced crime events [5]. These approaches have shown advantages in capturing spatial and temporal dependencies. However, they also often require large datasets, high geographic resolution, and complex training processes—conditions that are not always available in regional registries from Latin American countries [6]. Given these limitations, unsupervised learning methods offer a relevant alternative for exploring latent structures in crime databases without requiring prior labeling [7], [8]. Clustering allows grouping of territorial units based on similarities in their crime profiles, rather than solely on their positions in incidence rankings [8]. Among these methods, K-Medoids is particularly suitable when extreme values or regions with atypical behavior exist, because it uses real-world observations as representative centers of the groups [9]. The study's contribution is threefold. First, it proposes a regional-year matrix based on crime rates per 100,000 inhabitants, which enables comparison of territories with different population sizes and overcomes the limitations of absolute counts [9], [10]. Second, it applies K-Medoids to identify a granular partition of ten distinct territorial typologies of crime (k=10), prioritizing the multi-variable interpretability of the groups and the robustness of the algorithm in the face of extreme values or regions with atypical behavior [8], [9]. Third, it incorporates a spatial reconfiguration analysis by comparing profiles over a decade, which allows for the differentiation of persistent, emerging, declining, or volatile regions [1], [5]. With this, the article seeks to provide empirical evidence for the design of differentiated public policies, moving beyond the logic of uniform intervention based solely on national incidence rankings [10], [11]. Therefore, the objective of this research is to classify and characterize territorial crime patterns in Peru using unsupervised learning and K-Medoids [8], [9], employing regional rates per 100,000 inhabitants for the period 2015–2024 and validating the model's stability to ensure the consistency of the findings [1], [12]. This approach allows us to determine which regions share similar crime profiles, which types of crimes define each cluster, and which territories show signs of spatial reconfiguration during the period under analysis, thereby facilitating more efficient resource management [8], [10].

II. METHODOLOGY

1. Study design

This study adopts a quantitative, longitudinal, exploratory, and unsupervised learning design, aimed at classifying territorial crime patterns in Peru during the period 2015–2024 [1], [8]. The longitudinal approach enables analysis of the evolution of regional crime profiles over time [11], [13], while unsupervised learning identifies latent structures without imposing prior categories of risk or criminality [9]. This type of strategy is relevant to studies of territorial crime, since crimes are not distributed homogeneously but tend to concentrate in specific areas and periods depending on conditions of opportunity, mobility, institutional control, and risk exposure [1], [11], [12], [13]. The choice of an unsupervised model also stems from the limitations of purely predictive approaches when reliable labels or sufficient historical data are unavailable [9], or when the target variable may be biased by administrative records [6]. Therefore, this study does not seek to predict the aggregate number of crimes, but rather to classify territories using multivariate profiles constructed from specific rates for

each type of crime [8], [10]. This decision avoids tautological relationships between aggregate variables and their components. It shifts the focus from artificial prediction to the identification of interpretable territorial typologies, a relevant methodological requirement in applied criminological studies to guarantee transparency and accountability [5], [14].

2. Extraction, purification, and territorial homologation

The information downloaded from DATACRIM (https://datacrim.inei.gob.pe/panel/mapa) underwent a data cleaning and territorial standardization process before being incorporated into the machine learning model. The original tables were organized in a matrix format, with territorial units in rows and years in columns. To ensure territorial comparability, these tax units were standardized to the 25 regional units considered in the study: the 24 departments of Peru and the Constitutional Province of Callao. Population data from the INE (National Institute of Statistics and Informatics) were used to calculate the total population, estimated and projected as of June 30, by year, sex, and age group, 2000-2026 (https://www.inei.gob.pe/estadisticas/indice-tematico/poblacion-y-vivienda/).

3. Study period

The main analysis period spans the years 2015–2024. This time frame was selected because it provides comprehensive and comparable coverage of the crime categories included in the model and allows observation of the evolution of territorial patterns before, during, and after the COVID-19 pandemic. Choosing a ten-year time series strikes a balance among temporal breadth, data consistency, and statistical stability for the clustering analysis. As a validation strategy, the analysis is replicated for the period 2015–2021, enabling backtesting to confirm whether historical patterns remain valid or whether drastic changes have occurred that require recalibration [6], [15]. This validation allows us to assess whether the territorial profiles identified over the full 2015–2024 period are maintained in a more conservative time subset, thereby verifying whether the clusters are stable and persistent or represent transitory fluctuations [1], [13]. Furthermore, it allows us to compare the stability of the clusters using population estimates, ensuring that the findings capture a structural crime "signature" that is consistent across years [1], [11]. This logic of temporal validation is consistent with the spatiotemporal modeling literature, which recommends evaluating the stability and robustness of the patterns identified in different segments of the series [1], [15], [16].

4. Inclusion and exclusion criteria

For the main model (2015-2024), crime categories were selected that met five of DATACRIM's methodological criteria (such as regional coverage, complete annual data, rates per 100,000 inhabitants, and criminological relevance). These categories include crimes against property, against public administration, against the family, against liberty, against public safety, against public trust, against life, body, and health, reports of illicit drug trafficking, juvenile offenders, crimes against public order, domestic violence, and sexual violence against women. The year 2025 was excluded because only partial data (January-August) were available, thus avoiding unequal comparisons.

5. Transformation to long format for machine learning

This format organizes each record as a unique combination of region, crime, and year, with their respective values for recorded crimes, population, and rate. The general structure adopted was as follows:

$$REGION, CRIME, YEAR, CRIMES, POPULATION, RATE_100K$$

Subsequently, for the application of K-Medoids, the long-format database was transformed into a wide-format, where each row represents a region-year observation and each column corresponds to a crime category, expressed as a rate per 100,000 inhabitants. This transformation allows each region in each year to be represented as a multivariate vector of crime (equation 1):

$$X_{r,t} = (x_1, x_2, x_3, \ldots, x_p) \quad (1)$$

Here:$X_{r,t}$: represents the crime profile of the region in year t.

- $x_1, x_2, \ldots, x_p$: represent the rates per 100,000 inhabitants for each crime category.
- p: represents the total number of criminal variables included.

6. Final analytical matrix

The main matrix of the study comprises 25 regions and 10 years, resulting in a total of:

$$25 \times 10 = 250 \text{ region-year observations}$$

Each observation contains the rates per 100,000 inhabitants for the selected crime categories. This configuration allows for the simultaneous incorporation of territorial heterogeneity and the temporal evolution of crime. Unlike a static classification based on regional averages, the region-year unit allows observation of changes, stability, or reconfiguration in regional crime profiles over the period 2015–2024.

The validation matrix considers the period 2015–2021, generating:

$$25 \times 7 = 175 \text{ region-year observations}$$

This second matrix allows us to test whether the cluster structure obtained in the main model holds when the analysis is restricted to the most conservative time frame. Methodologically, this validation increases the credibility of the results and reduces the risk that the identified patterns depend exclusively on the 2022–2024 period.

7. Standardization of variables

Standardization was performed using z-scores; as a sensitivity analysis, robust standardization based on the median and interquartile range will also be evaluated (equation 2):

$$z_{i,j}^{rob} = \frac{x_{i,j} - Mediana(x_j)}{IQR(x_j)} \quad (2)$$

Where:

- i: represents the specific region-year observation.

- j: represents the specific crime category or variable.$z_{i,j}^{rob}$: represents the robust standardized value for the region-year observation i in crime category j.
- $i.,i,x_{i,j}$: represents the original crime rate per 100,000 inhabitants for the region-year observation i in crime category j.
- $Mediana(x_j)$represents the median rate of crime category j across all N observations.
- $IQR(x_j)$represents the interquartile range of crime category j, calculated as the difference between the 75th and 25th percentiles (Q3 - Q1).

This alternative is relevant when there are regions with extreme rates, given that territorial crime can exhibit asymmetric distributions, high spatial concentration, or outliers associated with specific territories. The literature on machine learning applied to crime data recognizes that spatial heterogeneity, territorial concentration, and the presence of sparse or extreme data pose significant challenges to building stable and interpretable models [14], [15].

8. Classification Algorithm: K-Medoids

The main algorithm used was K-Medoids, an unsupervised learning technique that groups observations according to their multivariate similarity. Unlike K-Means, which represents each cluster by its centroid, K-Medoids uses real observations as representative centers, or medoids [9]. This feature facilitates the substantive interpretation of the groups. It reduces sensitivity to extreme values, which is especially relevant in territorial crime databases where some regions may exhibit atypical rates for certain crimes, preventing these extreme cases from distorting the overall structure of the clusters [10], [11] (equation 2).

$$\min \sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, m_k) \tag{3}$$

Where:

- $The\ K\ represents\ the\ number\ of\ clusters.$: represents the set of observations assigned to the cluster k.
- $i.\bullet\ represents$ the observation.
- m_k: represents the medoid of the cluster.
- $d(x_i, m_k)$: represents the distance between the observation and the medoid.

9. Ethical and interpretative considerations

The study uses aggregated regional-level information and therefore does not involve personal data or individual identification. However, it is acknowledged that territorial classification models can pose interpretive risks if clusters are used as deterministic labels for danger, which could lead to unjustified over-patrolling in specific communities. Therefore, the results are interpreted as statistical profiles of recorded crime, not as normative

judgments about regions or populations, recognizing that official data capture only a fraction of the reality of crime.

10. Processing Reproducibility

To ensure reproducibility, the processing is documented in four stages: download from DATACRIM, territorial standardization, conversion to rates per 100,000 inhabitants, and transformation into a machine learning matrix. Each stage maintains traceability between the source and the final matrix used in the model.

The methodological flow can be summarized as follows:

$$Datos\ DATACRIM \rightarrow Homologación\ territorial \rightarrow Conversión\ a\ tasas\ 100K \rightarrow Formato\ largo \rightarrow Formato\ ancho \rightarrow Estandarización \rightarrow K\text{-}Medoids$$

III. RESULTS

1. Evolution of crime rates

The evolution of crime rates shows that crime in Peru did not follow a uniform trajectory between 2015 and 2024. Instead, differentiated trends were observed across crime categories. The greatest relative increase was observed in crimes against life, body, and health, whose regional average rate rose from 255.38 to 1,154.44 per 100,000 inhabitants, representing a 352.05% increase. This result demonstrates a substantial reconfiguration of the recorded crime profile, particularly in crimes associated with direct harm to personal integrity. Likewise, significant increases were recorded in crimes against the family (87.89%), juvenile offenders (75.94%), sexual violence against women (71.54%), and crimes against property (59.51%), as detailed in Table 1 and Figure 1. More moderate increases were observed in crimes against liberty (41.30%), public safety (33.37%), public administration (25.04%), illicit drug trafficking (9.99%), and public trust (2.53%). In contrast, crimes against public order (-13.03%) and domestic violence (-98.00%) showed declining average rates over the study period. These findings are consistent with the literature indicating that crime patterns vary significantly according to the type of crime, the specific criminal opportunities of each location, territorial structure, and institutional conditions [1], [11], [12], [17].

Table 1. Evolution of cases and average rates by type of crime, 2015–2024.

Crime	Cases 2015	Average rate 2015	Cases 2024	Average rate 2024	Growth in cases (%)	Average growth rate (%)
against liberty	40969	151.4	68905	213.93	68.19	41.3
against life, body, and health	64490	255.38	375779	1154.44	482.69	352.05
against property	188869	714.45	405846	1139.63	114.88	59.51
against public administration	51000	205.15	74867	256.51	46.8	25.04
against public order	4547	10.71	3150	9.31	-30.72	-13.03
against public safety	71053	266.11	105674	354.92	48.73	33.37
against public trust	23737	75.18	25529	77.08	7.55	2.53
against the family	45895	194.96	97240	366.32	111.87	87.89
domestic violence	137181	527.92	1964	10.58	-98.57	-98.0
illicit drug trafficking	14453	39.91	15882	43.9	9.89	9.99

juvenile offenders	13557	47.8	25375	84.09	87.17	75.94
sexual violence against women	5311	21.51	10111	36.9	90.38	71.54

Note: Average rates are the annual regional averages per 100,000 inhabitants.

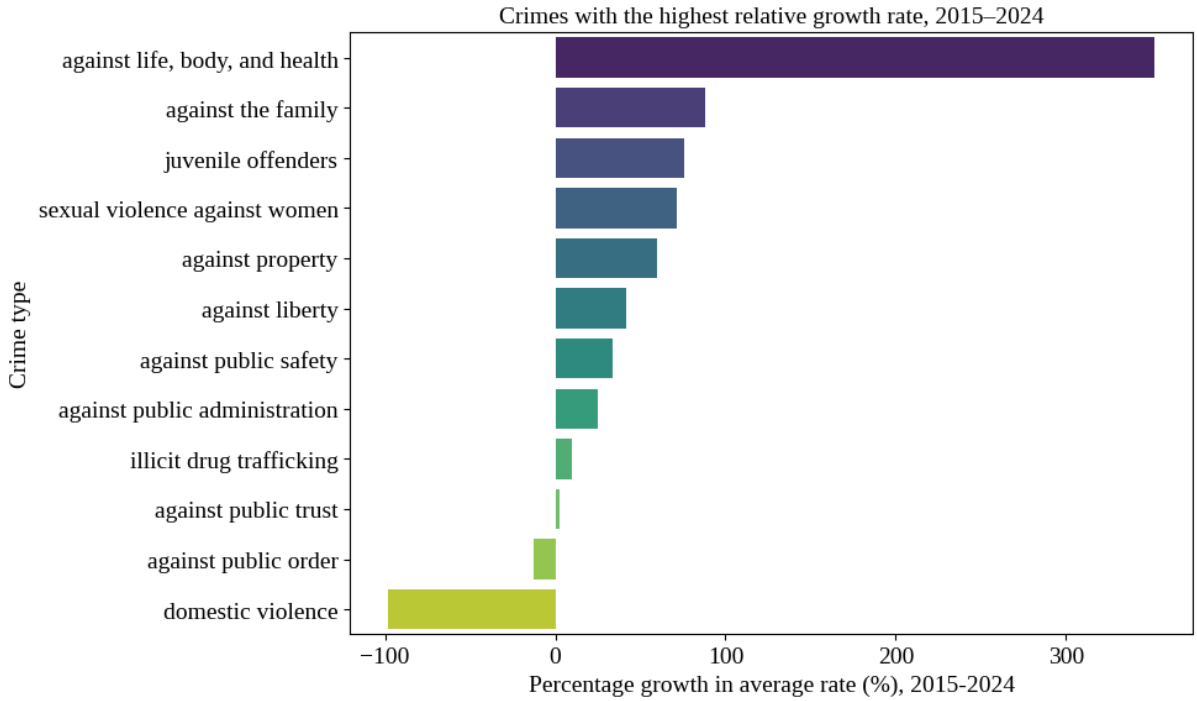


Figure1. Crimes with the highest relative growth rate, 2015–2024.

The figure shows that the largest relative increases correspond to crimes against life, body, and health, crimes against the family, juvenile offenders, sexual violence against women, and crimes against property.

2. Selection of the optimal number of clusters

To determine the optimal partition, solutions with k = 2 to 10 were evaluated using multiple internal validation metrics. The results identified k = 10 as the optimal clustering solution, as it achieved the lowest Davies–Bouldin index, the highest Calinski–Harabasz index, the largest Gap Statistic, and the best overall average ranking (Table 2, Figure 2).

Table 2. Internal validation for the selection of the number of clusters.

k	Manhattan Cost	Silhouette	Davies-Bouldin	Calinski-Harabasz	Gap Statistic	Average ranking
2	2171.4022	0.1112	3.2566	7.9006	1.0822	3.2
3	2044.772	0.0002	4.1628	5.5264	1.0271	4.0
4	2014.6868	-0.0283	4.5453	4.6123	0.9945	5.4

5	1950.832	-0.0711	5.6668	4.3017	0.9856	6.2
6	1940.8765	-0.0904	5.4279	3.4839	0.9538	7.0
7	1871.319	-0.1454	5.8565	3.5531	0.9632	7.0
8	1843.265	-0.1805	4.3412	5.2538	0.9685	5.0
9	1861.6794	-0.206	3.7777	5.1684	0.9391	5.8
10	1560.0176	-0.0202	2.6714	74.3815	1.0963	1.4

Note: The optimal solution was selected at k = 10, as it achieved the best overall performance across the internal validation metrics, including the lowest Davies–Bouldin index, the highest Calinski–Harabasz index, the largest Gap Statistic, and the best average ranking.

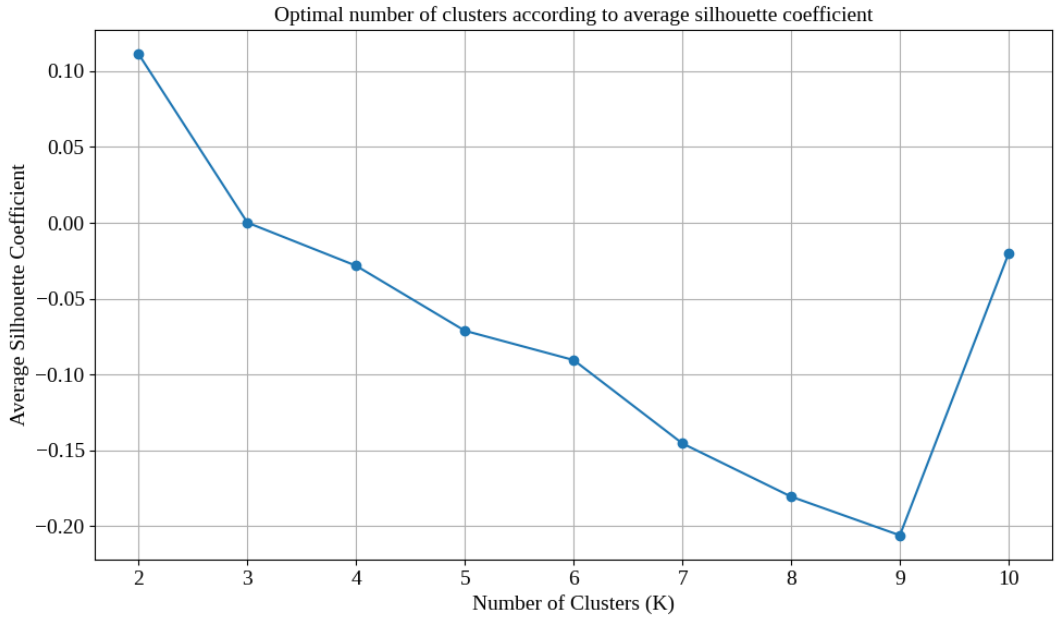


Figure 2. Optimal number of clusters according to the average silhouette coefficient.

The variation of the internal validation metrics (Figure 2) indicates that k = 10 provides the best overall clustering performance, supporting its selection as the final partition.

It is critical to note that although the silhouette coefficient attains its highest value at k = 2 (Silhouette = 0.1112), this partition was discarded due to its limited practical utility and excessive analytical overgeneralization in criminological analysis. When evaluating the joint behavior of the remaining internal validation metrics, the k = 10 solution achieves the optimal global performance: it records the lowest Davies–Bouldin index (2.6714), a critical and robust increase in the Calinski–Harabasz index (74.3815), and the largest Gap Statistic (1.0963). This multi-criteria consensus mathematically justifies selecting a granular partition of 10 territorial crime profiles.

3. Main classification with K-Medoids

The final K-Medoids solution (k = 10) identified ten distinct territorial crime profiles across the region-year observations. Cluster sizes ranged from 7 to 54 observations and involved between 4 and 18 regions, indicating

that the identified clusters represent dynamic region-year crime profiles rather than fixed regional groupings. This classification captures the temporal and spatial heterogeneity of recorded crime patterns across Peru. The dominant crime composition varied substantially across clusters (Table 3 and Figure 3). While some clusters were characterized by higher standardized averages of illicit drug trafficking, domestic violence, and crimes against public order, others were dominated by crimes against public safety, crimes against property, juvenile offenders, crimes against public administration, or crimes against life, body, and health. These differences reveal heterogeneous crime profiles, highlighting the coexistence of multiple territorial patterns rather than a single nationwide crime structure.

Table 3. Summary of the main classification with K-Medoids, k = 10.

Cluster	**Region-year observations**	**Regions involved**	**Dominant crimes according to the standardized average**
1	16	6	illicit drug trafficking; against public order; against liberty; against public safety
2	20	8	domestic violence; juvenile offenders; sexual violence against women; against public order
3	13	5	domestic violence; against property; illicit drug trafficking; against the family
4	18	9	illicit drug trafficking; against public trust; against public safety; domestic violence
5	24	13	domestic violence; illicit drug trafficking; sexual violence against women; against public safety
6	7	4	illicit drug trafficking; juvenile offenders; against public safety; against the family
7	42	7	domestic violence; sexual violence against women; against public order; against the family
8	19	18	domestic violence; against public order; against public trust; illicit drug trafficking
9	54	13	against public safety; against property; juvenile offenders; against public order
10	37	10	against public administration; domestic violence; against public trust; against life, body, and health

Note: The dominant crimes correspond to the variables with the highest standardized means within each cluster.



Figure 3. Average criminal profile by cluster.

The heat map visualizes the relative intensity of each crime type within the clusters, expressed as a robust standardized mean.

4. Longitudinal validation 2015–2021

To assess the model's temporal stability, the clustering procedure was replicated for the period 2015–2021. This validation allowed comparison of whether the crime profiles identified in the main period (2015–2024) were preserved when the analysis was restricted to a shorter time frame. The validation again selected k = 10 as the optimal solution, indicating consistency in the identified clustering structure. The agreement between the partitions was assessed using the Adjusted Rand Index (ARI), yielding a value of 0.6046 based on 175 comparative observations. This result indicates a moderate level of agreement between the main and validation solutions, suggesting that the identified crime profiles remain relatively stable over time despite temporal variations. Methodologically, these findings support the robustness of the clustering model, demonstrating that the classification is not exclusively dependent on the selected observation period but reflects persistent spatial patterns of recorded crime [1], [11], [13]. Longitudinal validation (Table 4) is particularly relevant in spatiotemporal studies, where crime patterns may vary according to the observation period, spatial scale, and the dynamics of administrative records [1], [9], [11], [15].

Table 4. Longitudinal validation of the K-Medoids model.

Model	Period	Optimal k	Comparative observations	ARI
Main model	2015-2024	10	175	0.604614
Validation model	2015-2021	10	175	0.604614

Note: The ARI measures the agreement between partitions adjusted for chance. Values close to 1 indicate greater stability between solutions.

5. Territorial reconfiguration

The transition analysis between 2015 and 2024 allowed us to assess whether regions maintained their membership in the same crime profile or shifted clusters over the study period. The transition matrix reveals movements across the ten identified profiles, confirming that recorded crime should not be understood as a fixed territorial condition, but rather as a dynamic process of spatial and temporal reconfiguration [5], [11]. This approach complements the static classification with a longitudinal perspective of regional mobility, consistent with previous research on crime displacement, concentration, and territorial change [1], [17]. The transition matrix between clusters (Figure 3) is particularly useful for identifying persistent regions, emerging profiles, and territories undergoing changes in their relative crime structure [13]. This approach moves beyond isolated rankings, as a region may not consistently occupy the highest positions in absolute crime levels but may experience substantial changes in its crime composition or temporal pattern [11]. From a public policy perspective, this analysis helps distinguish territories with persistent crime profiles from those undergoing transformation, supporting more targeted resource allocation and decision-making adapted to the dynamic evolution of crime patterns [18].

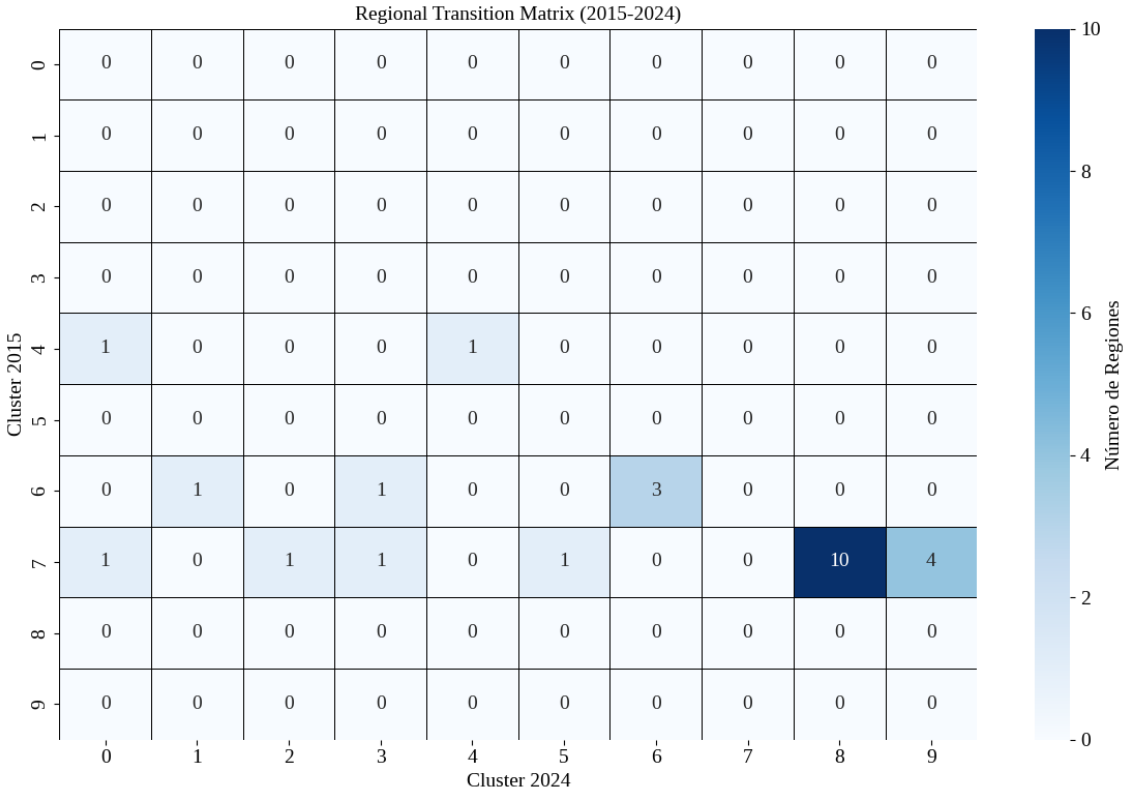


Figure 4. Regional transition matrix between clusters, 2015–2024.

The figure shows changes in cluster membership among Peruvian regions between 2015 and 2024. The matrix values indicate the number of regions that remain within the same cluster or transition between the ten identified territorial crime profiles.

6. Summary of results

Taken together, the results show that recorded crime in Peru during 2015–2024 exhibits a heterogeneous territorial structure. The primary solution using K-Medoids with k = 10 identified ten distinct territorial crime profiles, supported by multiple internal validation metrics and reinforced through longitudinal validation, which showed moderate agreement between the main and validation partitions (ARI = 0.6046).

These findings fulfill the study's objective by demonstrating that unsupervised learning enables the classification of territorial crime patterns without relying on aggregated dependent variables or tautological models. Furthermore, using rates per 100,000 inhabitants and temporally validating clusters strengthens the methodological robustness of the analysis, indicating that the identified profiles reflect persistent criminal patterns rather than random fluctuations. Additionally, incorporating territorial transition analysis highlights crime as a dynamic phenomenon shaped by spatial and temporal changes. Overall, this approach goes beyond descriptive and fragmented interpretations of criminality by providing a multivariate characterization of regional crime profiles, integrating diverse dimensions of crime to support more precise, evidence-based public policies.

IV. DISCUSSION

The multidimensional crime structure identified in Peru (2015–2024) confirms that territorial crime is not randomly distributed, but rather responds to differentiated socio-spatial dynamics. The selection of the optimal partition of k=10 clusters, supported by the best overall performance in internal validation metrics (lowest Davies–Bouldin index and highest Calinski–Harabasz index), demonstrates that the country's crime reality requires granular segmentation to avoid masking critical patterns. The silhouette coefficient for this solution reflects the inherent complexity of the social data, in which regional boundaries naturally overlap due to territorial interconnections and demographic variability. The model's temporal stability is one of its most robust methodological components. The Adjusted Rand Index (ARI = 0.6046) from the longitudinal validation (2015–2021) indicates a moderate and statistically significant level of agreement between the partitions. This ensures that the clustering with k=10 identifies structural and persistent criminal signatures over time, rather than responding to random annual fluctuations or artifacts derived from population projections. Detailed multivariate analysis isolated profiles with high criminological specificity. For example, Cluster 9 shows a critical concentration of crimes against public safety and property, as well as juvenile offenses. At the same time, Clusters 2, 5, 7, and 10 demonstrate the persistence of domestic and gender-based violence dynamics at varying relative intensities. This overview provides direct practical value for public safety planning by overcoming the ineffectiveness of absolute administrative rankings. It thus facilitates the transition from uniform interventions to differentiated strategies: situational and targeted policing for property-related profiles, and intersectoral coordination (health and justice) in regions dominated by interpersonal violence. Finally, the shifts recorded by the transition matrix corroborate the view that criminal behavior is not a static geographical condition but rather a dynamic process of spatial reconfiguration. Although the model is limited by the "dark figure" of unreported crimes, the methodological traceability based on DATACRIM-INEI records establishes a reproducible standard for regional criminological analysis.

V. LIMITATIONS

The study has methodological limitations that must be considered when interpreting the findings. Firstly, the data analyzed correspond exclusively to the crime recorded administratively by DATACRIM–INEI, which implies that the results may be influenced by under-reporting or the "black figure" of crime, regional variations in the propensity to report, and disparities in institutional recording capacities. Secondly, the regional aggregation scale, although it guarantees comparability at the national level, tends to mask heterogeneity at the provincial, district, or intra-urban level. Thirdly, the initial omission of specific categories such as homicide restricts the analysis of lethal violence, which has logics of social gravity and spatial dynamics different from property crimes. Finally, the K-Medoids algorithm classifies multivariate profiles and temporal reconfigurations but does not establish direct causal relationships. Future developments must incorporate socioeconomic variables (poverty, inequality, urbanization) to deepen the model's predictive and preventive scope.

VI. CONCLUSIONS

This study successfully classified and characterized territorial crime patterns in Peru (2015–2024) using normalized rates per 100,000 inhabitants and unsupervised learning with K-Medoids. The results demonstrate that crime exhibits a heterogeneous and dynamic spatial structure that goes beyond the linear interpretations of traditional rankings. The evolution of rates revealed a profound reconfiguration of the crime landscape, highlighting the asymmetric relative growth of crimes against life, body, and health (+352.05%), followed by substantial increases in crimes against the family and juvenile offenders. The modeling indicated that the partition with k=10 clusters is the statistically most appropriate solution for describing the complexity of the

Peruvian territory. The stability of this structure was corroborated by longitudinal validation from 2015 to 2021, yielding an ARI of 0.6046, confirming consistent structural patterns despite variations in the observation period. The resulting taxonomy allowed for the isolation of specific criminal dynamics (such as drug markets, crimes against public administration, or interpersonal violence), providing a detailed analytical map of the coexistence of multiple regional criminal profiles. The central contribution of this research is a parsimonious, reproducible, and ethically aligned methodological framework that promotes transparency in the use of official data for evidence-based decision-making. From an applied perspective, the findings indicate that security policies in Peru must abandon uniform patrolling strategies and shift toward targeted, multidimensional interventions: situational crime prevention in areas with high property values and intersectoral coordination between the justice and health sectors in regions vulnerable to domestic and gender-based violence.

REFERENCES

[1]. F. Denegri y J. Ley-García, «Crime against Businesses: Temporal Stability of Hot Spots in Mexicali, Mexico», *ISPRS Int. J. Geo-Inf.*, vol. 10, n.º 3, p. 178, mar. 2021, doi: 10.3390/ijgi10030178.

[2]. K. Kim y Y.-A. Kim, «A Machine Learning Approach to Analyzing Crime Concentration: The Case of New York City», *Justice Q.*, vol. 42, n.º 7, pp. 1300-1320, Nov. 2025, doi: 10.1080/07418825.2024.2401938.

[3]. D. Bonilla-Calle, E. D. Nascimento, and M. M. V. Arias, «Pandemic, Criminal Governance and Mechanisms of Social and Economic Regulation in Three Latin American Contexts», *Dilemas Rev. Estud. Conflito E Controle Soc.*, vol. 15, n.˚ Especial 4, pp. 287-310, 2022, doi: 10.4322/dilemas.v15esp4.52508.

[4]. R. Romero-Carazas *et al.*, «Urban Victimization in Peruvian Cities: Influence of Demographic and Population Factors on Citizen Security», *Qubahan Acad. J.*, vol. 6, n.º 1, pp. 194-210, ene. 2026, doi: 10.48161/qaj.v6n1a2101.

[5]. Y. Du y N. Ding, «A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods», *ISPRS Int. J. Geo-Inf.*, vol. 12, n.º 6, p. 209, may 2023, doi: 10.3390/ijgi12060209.

[6]. J. D. Gelvez Ferreira, M.-P. Nieto-Rodríguez y C.-A. Rocha-Ruiz, «Prediciendo el crimen en ciudades intermedias: un modelo de "machine learning" en Bucaramanga, Colombia», *URVIO Rev. Latinoam. Estud. Segur*, n.˚ 34, pp. 83-98, sep. 2022, doi: 10.17141/urvio.34.2022.5395.

[7]. U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, y H. H. R. Sherazi, «Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review», *IEEE Access*, vol. 8, pp. 166553-166574, 2020, doi: 10.1109/ACCESS.2020.3022808.

[8]. T. J. Fontalvo-Herrera, M. A. Vega-Hernández y F. Mejía-Zambrano, «Método de clustering e inteligencia artificial para clasificar y proyectar delitos violentos en Colombia», Rev. Científica Gen. José María Córdova, vol. 21, n.º 42, pp. 551-572, abr. 2023, doi: 10.21830/19006586.1117.

[9]. F. Díaz, N. Cerna, y R. Liza, «Artificial Intelligence and Crime in Latin America: A Multilingual Bibliometric Review (2010–2025)», *Information*, vol. 16, n.º 11, p. 1001, nov. 2025, doi: 10.3390/info16111001.

[10]. J. C. D. O. R. Costa y M. M. Silva, «A CLUSTERING-BASED APPROACH FOR IDENTIFYING GROUPS OF MUNICIPALITIES TO SUPPORT THE DIRECTION OF PUBLIC SECURITY POLICIES», *Pesqui. Oper.*, vol. 42, p. e257930, 2022, doi: 10.1590/0101-7438.2022.042.00257930.

[11]. R. Prieto Curiel, «Weekly Crime Concentration», *J. Quant. Criminol.*, vol. 39, n.º 1, pp. 97-124, mar. 2023, doi: 10.1007/s10940-021-09533-6.

[12]. N. T. Connealy, «The Influence, Saliency, and Consistency of Environmental Crime Predictors: A Probability Score Matching Approach to Test What Makes a Hot Spot Hot», *Justice Q.*, vol. 40, n.º 5, pp. 670-693, jul. 2023, doi: 10.1080/07418825.2022.2106292.

[13]. V. Harinam, Z. Bavcevic, and B. Ariel, «Spatial distribution and developmental trajectories of crime versus crime severity: do not abandon the count-based model just yet», Crime Sci., vol. 11, n.o 1, p. 14, Nov. 2022, doi: 10.1186/s40163-022-00176-x.
[14]. A. P. Wheeler and W. Steenbeek, «Mapping the Risk Terrain for Crime Using Machine Learning», *J. Quant. Criminol.*, vol. 37, n.º 2, pp. 445-480, jun. 2021, doi: 10.1007/s10940-020-09457-7.
[15]. C. Jing *et al.*, «A deep multi-scale neural network for crime hotspot mapping prediction», *Comput. Environ. Urban Syst.*, vol. 109, p. 102089, abr. 2024, doi: 10.1016/j.compenvurbsys.2024.102089.
[16]. Y. Sun, T. Chen, and H. Yin, «Spatial-temporal meta-path guided explainable crime prediction», *World Wide Web*, vol. 26, n.º 4, pp. 2237-2263, jul. 2023, doi: 10.1007/s11280-023-01137-3.
[17]. R. Massa y G. Fondevila, «Criminal displacement in Mexico City's metropolitan area: The case of kidnapping», *Int. J. Law Crime Justice*, vol. 67, p. 100479, Dec. 2021, doi: 10.1016/j.ijlcj.2021.100479.
[18]. D. T. O'Brien, A. Ristea, F. Hangen, and R. Tucker, «Different places, different problems: profiles of crime and disorder at residential parcels», Crime Sci., vol. 11, n.o 1, p. 4, Dec. 2022, doi: 10.1186/s40163-022-00165-0.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Optimized Algorithm towards Addressing the Cold-Starts Issue in the Airline Recommender System

Ayman Mahmud
Department of Computer Science, Faculty of Science and Natural Resources, Aljafara University, Libya,
Ayman@aju.edu.ly

Zakaria Suliman Zubi
Department of Computer Science, Faculty of Information Technology, Sirte University, Sirte, Libya,
zszubi@su.edu.ly

Asma Alhadi Billayl
Department of Computer Science, Faculty of Science and Natural Resources, Aljafara University, Libya,
Asmabill79@gmail.com

Adud Musa Mabu
Department of Computer Science, Yobe State University, Damaturu, Nigeria
audumusamabu@ysu.edu.ng

Abstract
Recommender systems are techniques created to provide tailored suggestions for airlines and their ancillaries that are most pertinent to each passenger. The Passenger Cold Start Problem occurs when a new passenger cannot be properly recommended because there is insufficient historical data to suggest airlines or ancillaries that align with the passenger's interests. The proposed techniques efficiently personalize recommendations for each passenger using available historical data to satisfy each passenger's experience. The dataset comprises three airline carrier datasets detailing every commercial flight within Libya, the UAE, and Turkey. The Python platform was used to actualize the algorithm. Initial flight recommendations to new passengers are made while the model continuously updates their profiles based on their interactions with the airline booking engine, thereby enabling continuous learning. As more data becomes available, the algorithm refines its recommendations, leveraging ongoing interactions to optimize the accuracy using the Bipartite Network Algorithm (BNA). The Pearson correlation coefficient was used to compute the passenger-Airline matrix of passenger preferences. $\mathbf{W_u}$), and the minimum similarity threshold is set to η to determine the similarity. The performance of the proposed model was evaluated using Mean Absolute Error (MAE) and Precision, which measured the proximity of predicted grades to true ratings. The results show that for new passengers, the Emirates Airlines data has an MAE of 1.248 for SVD, 0.947 for Bipartite-Cluster, and 0.916 for the proposed algorithm (BNA). Turkish Airlines data has an MAE of 1.319 for SVD, 1.291 for Bipartite-Cluster, and 0.946 for the proposed algorithm, and Afriqiyah Airways has an MAE of 1.319 for SVD, 1.291 for Bipartite-Cluster, and 0.964 for the proposed algorithm. The result clearly indicates that the Optimized Bipartite Algorithm outperformed SVD and Bipartite-Cluster.

Keywords: Recommender System, Cold Start problem, Bipartite Algorithm, Transfer Learning, sentiment Analysis, Mean Absolute Error, Execution Time, passenger, Pearson correlation coefficient, and Airlines.

I. INTRODUCTION

Sentiment analysis has become highly valuable in various fields, including e-commerce, customer relationship management, and social media. By providing insights into users' opinions, attitudes, and feelings, emotion analysis could significantly improve the reliability of recommendation techniques. Integrating emotion analysis into these techniques implies more precise, tailored recommendations, eventually enhancing user satisfaction and their interaction. Recommender techniques can be implemented using two

methods: collaborative filtering and the content-based approach. Many systems rely on a hybrid of these methods.
Content-based approaches assume that if a user were concerned with an item in their last experience, they would be concerned with the same item in the next experience as well. By building a user account based on their historical attitude (actions) and gathering identical items by their properties, the engine will suggest new items to the user. Furthermore, Collaborative filtering aims to provide intelligent recommendations based on past user experience, represented as an interaction matrix between users and items. This method could be described as a generalization of classification and regression issues in which the features cannot be distinguished from the targets. i.e., the algorithm examines to predict the value of each cell in the said matrix.

Sentiment analysis of user ratings and attitudes has become an important basis for recommender systems [1], enabling them to predict users' future interests based on their experience. Taken together, this improves the system's ability to suggest accurate recommendations to users individually. Significant challenges face recommender techniques, especially the collaborative filtering approach [1], which suffers from the cold-start issue for new users or items. Totally, content-based recommender techniques usually lack variety in their suggestions and recommendations, which implies lower system accuracy. Handling the said challenge could enhance the performance of sentiment-based recommender techniques. Integrating methods such as deep learning and hybrid approaches can significantly enhance performance. A recommendation system can help users quickly and accurately find information that meets their needs [2]. This improves user satisfaction and user experience [3]. Today, recommendation techniques play a crucial role in e-commerce [4], [5] and other domains [6-11], [3], [12-15]. Moreover, Recommender techniques play a crucial role by providing users with services tailored to their preferences.

In this research, we examined the most important challenges in recommendation techniques. We address the cold-start issue for new passengers by merging sentiment analysis from existing passenger ratings. This method helps us determine what is interesting to new passengers and suggest an airline with the highest ratings, even if the new passenger's required data is still unknown. However, the cold-start issue is common in recommender techniques [16], [17]. Table 1 elaborates on the cold-start issue and clarifies it by providing an example of a recommendation based on passenger ratings. In Table 1, rows represent passengers. The columns represent n airlines. The matrix shows passenger reviews of the airlines in the range {1, 2, 3, 4, 5}, where the reviews are integers. Values 0-5 represent passengers' reviews of the airline, and if a passenger has not reviewed the airline, it is denoted by '?'.

Table 1. Illustration Sample of cold-start problem in the passenger-airline rating matrix

Passenger/ Airline	**Airline 1**	**Airline 2**	**Airline 3**	**.....**	**Airline n**
Passenger 1	3	4	?	.....	?
Passenger 2	?	5	4	……	?
Passenger 3	1	0	4	……	?

II. LITERATURE REVIEW

Several studies on the cold-start issue for new users are presented in the following paragraphs. Panda and Ray [18] performed research and a review of ninety-one papers on cold-start mitigation. The research aimed to categorize the several methods into data-driven and method-driven techniques. It also classified method-driven techniques into five main methods depending on the models used by the scientists. The research also

emphasized dozens of algorithms for comparison, and some of the metrics used were analyzed to evaluate the cold-start issue. However, the study was unable to distinguish user cold starts from cold starts in the project. Since there is no relation between these two issues, the solutions to each problem should be determined separately.

Son [19] also conducted an important study on the cold-start issue for new users. The examiners introduced a study aimed to analyze the compression of three methods for the problem, involving (i) utilizing some available information from other sources, (ii) selecting the foremost matching user classifications, and (iii) using hybrid methods to improve prediction. He presented a comparative study of several algorithms, including HU-FCF, MIPFGWC-CS, and FARAMS. Nevertheless, his systematic study was completed in 2014, and many statistical techniques that it didn't address have been developed recently.

Abdullah et al. [20] have offered a detailed categorization and an introduction to models to identify the new-user issue through promoted information. The study presented a comprehensive explanation of how supporting data can be acquired and extracted. The systematic study introduces a data-driven solution to the new-user cold-start problem. On the other hand, the study has given less focus to the method-driven technique. In short, their study has some issues with how the user cold-start problem could be solved. (i) The Issue is figured out in dated Literature's evaluations. Their analysis to figure out the Issue focuses entirely on one particular model. They do not consider other methods such as ML or DL.

H. Yuan et al. [21] stated in their article that the "user cold start problem" occurs when a new user cannot be effectively recommended due to a lack of historical information. This issue could obstruct the use of the recommendation system.

Their study covered all literature reviews conducted and published by candidates from January 2016 to April 2023, spanning 8 years. This study analyzed the literature on models, techniques, and approaches deployed to address the issue over the last eight years and classified them into two categories: Data-Driven Approaches and model-driven techniques. The study analyzed the existing evaluation criteria for these techniques to provide robust references for researchers in related areas.

Ke Yin Cao et al.[22] proposed a community detection algorithm for a cold start issue. Using the bipartite network, the similarity in this work has been calculated between the user and their preferred item. Community detection is performed using the Louvain algorithm on the projected One-Mode network.

Z. Zhang et al.[23] The proposed technique uses a link prediction method to simulate user identification and their preferences, reducing the range of possible recommendations. In addition, a Convolutional Neural Network (CNN) is used to predict the likelihood of flight cancellations or delays, enabling the model to provide passengers with recommendations that align with their preferences based on available information and flight details. The proposed model used clustering to simplify the handling of massive flight datasets and speed up processing. In their technique, the K-Means algorithm partitions the data into clusters, and the recommendation process is based on the cluster closest to the user's features. The outcome of the proposed technique was tested using actual flight data.

Aditya Naidu et al.[24] Presented a hybrid recommender system that combines content-based filtering, collaborative filtering, and sentiment analysis to give more accurate and user-centric tourism recommendations. By analyzing both item attributes and user interactions and incorporating feedback from user ratings, the model improves its predictions to better meet user needs. Their model is designed to eliminate information overload and optimize the efficiency and satisfaction of the journey-planning process. Their paper also evaluated the effectiveness of variety recommendation techniques and emphasized the Advantages and trade-offs connected with integrating hybrid models in travel techniques.

Karthiyayini, J [25] has proposed using multi-criteria recommendation techniques that consider multiple attributes, rather than relying solely on user ratings as in traditional recommendation systems. This strategy is based on a hybrid filtering technique. Furthermore, the model uses two-way automatic decoding (BDA) to personalize recommendations based on different user profiles. The introduced technique was tested using

several real-world datasets gathered from diverse travel platforms. The pilot evaluations showed its superior effectiveness at enhancing prediction accuracy compared with existing techniques. The proposed model not only addresses the challenges associated with large and diverse datasets but also provides personalized travel plans tailored to each user's preferences, thereby optimizing the user experience in travel recommendation systems.

III. RESEARCH METHODOLOGY

1. Transfer-learning-enabled BNCD for a new passenger

The proposed BNCD algorithm is trained in two phases. In the first phase, the model is trained on historical data from existing passengers. In this phase, BNCD learns a shared representation of passenger preferences and evaluates passenger-specific latent parameters. In the second phase, when a new passenger is introduced, the knowledge learned from existing passengers is transferred to the new passenger through three mechanisms:

1. Shared model initialization: the global BNCD parameters learned from existing passengers are utilized to initialize the model for the new passenger.
2. Bayesian prior transfer: the posterior distribution learned from existing passengers is used as the prior distribution for the new passenger.
3. Similarity-based passenger adaptation: the new passenger is compared to the majority of similar existing passengers or passenger clusters, and their learned parameters are utilized to initialize or systematize the new passenger model.

Transfer learning is embedded directly into the BNCD training and inference procedures.

1. Inputs

   Let there be *N* existing passengers. For passenger i, the Historical dataset is given by equation (1).

$$D_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i},$$

   Where $B\ddot{u}y\ddot{u}k$put feature vector aand $\text{yndy}_{i,j}$ points to the corresponding target, as shown in equation (2).$D_S = \{D_1, \dots, D_2\}$

   For a new passenger, the available data may be one of the following:

   1. No data, i.e., a cold-start scenario demonstrated in equation (3):

$$D_t = \emptyset$$

   2. A small support set illustrated in equation (4):

$$D_t^{sup} = \{(x_{t,j}, y_{t,j})\}_{j=1}^{n_t},$$

   Where $n_t It\ It\ It\ It$ is often small.

Optionally, side information for the new passenger may also be available, such as demographic attributes, travel preferences, historical route information, or other profile features. Let this information be denoted by

$$a_t.$$

Therefore, the complete inputs to the proposed algorithm are mentioned in equation (5):-

$$\{D_S\ , D_t^{sup}, a_t.\}$$

2. Model parameters

The BNCD model contains two types of parameters.

2.1 Global parameters

The global parameters are shared across passengers and are denoted by:

$$\theta.$$

The parameters, such as general behavioral structure, feature representations, or common latent factors, are called model common patterns and are shared by variant passengers

2.2 Passenger-specific parameters

For the passenger, the passenger-specific parameters are denoted by

$$\theta_i$$

These parameters capture individual differences among passengers. For the new passenger t, the corresponding passenger-specific parameters are:

$$\theta_t$$

The full parameter set for the passenger is therefore

$$\Theta_i = \{\theta, \phi_i\}.$$

3. Source training process

Firstly, BNCD is trained utilized the source dataset. The aim is to learn the shared parameters, the passenger-specific parameters, and the posterior distribution over them. The source training function is shown in equation (6):

$$L_S(\theta, \{\varphi_i\}) = \sum_{i=1}^{N} \sum_{(x,y)\in D_i} l\left(f_{\theta,\varphi}(x), y\right) + \lambda_{reg} R(\theta, \{\varphi_i\}) + \lambda_{KL} KL(q_S(\theta, \{\varphi_i\}) \| p_0(\theta, \{\varphi_i\}))$$

After source training, the model obtains as illustrated in equation (7):

$$\theta_S, \{\varphi_i\}_{i=1}^{N}, q_S(\theta, \{\varphi_i\}).$$

In addition, the algorithm stores transfer statistics, such as:

- passenger embedding or latent representations;
- passenger cluster assignments;
- cluster centroids;
- posterior means and covariances;
- Sufficient statistics required for Bayesian updating.

The knowledge is being transferred to a new passenger using the stored quantities mentioned.

3. Transfer learning mechanism

Once a new passenger t appears, the algorithm does not train from scratch. Instead, the information learned from existing passengers is transferred to the new passenger.

The transfer mechanism is defined as follows.

First, the source posterior is used to construct an initial target prior for the new passenger as demonstrated in equation (8):

$$q_t^0(\theta, \varphi_t) = T(q_S(\theta, \{\varphi_i\}))$$

Where T (.) is the transfer operator.

In the simplest case,

$$q_t^0(\theta, \varphi_t) = q_S(\theta, \{\varphi_i\}))$$

Meaning that the new passenger is initialized using the global source posterior.

In more personalized details, the algorithm calculates the similarity between the new passenger and each existing passenger by using equation (9):

$$w_i = \frac{\exp(s(a_t, a_i)/\tau)}{\sum_{k=1}^{N} \exp(s(a_t, a_k)/\tau)},$$

The transferred prior for the new passenger is then computed as a similarity-weighted combination of source-passenger posteriors as shown in equation (10):q alt t üssü 0 , sol parantez teta ,fi alt t, sağ parantez eşittir toplam alt limit i eşittir 1 üst limit N of w alt i, q alt i , sol parantez teta ,fi alt $i(\theta, \phi_t) = \sum_{i=1}^{N} w_i\, q_i(\theta, \phi_i)$.

Equivalently, if BNCD uses passenger clusters, the new passenger is assigned to the most relevant cluster, and the cluster posterior is used as the initialization.

Thus, the transferred information includes:

1. The global BNCD parameters ;
2. The posterior distribution learned from existing passengers;
3. Passenger embeddings or cluster centroids;
4. Similarity weights identifying the most relevant existing passengers;
5. Sufficient statistics required for fast Bayesian updating.
6. Adaptation for a new passenger

Once the transfer is prior q_t^0 Once it is obtained, the model adapts to the new passenger using any available information.

Data D_t^{Sup} .

The adaptation objective is illustrated in equation (11):

$$L_t(\theta, \phi_t) = \sum_{(x,y)\in D_t^{sup}} \ell\left(f_{\theta,\phi_t}(x), y\right) + \beta\, KL\left(q_t(\theta, \phi_t) \,\|\, q_t^0(\theta, \phi_t)\right) + \lambda_t \|\phi_t - \phi_t^0\|^2,$$

The adapted posterior is obtained by minimizing the above function, as shown in equation (12): q_altin equation (12):$q_t^* = \underset{q}{\operatorname{argmin}} L_t(\theta, \phi_t)$.

If only a very small amount of data is available, the global parameters may be frozen, and only the Passenger-specific parameters are updated:$\theta = \theta_S \phi_t = \underset{\phi_t}{\operatorname{argmin}} L_t(\theta_S, \phi_t)$.

This prevents overfitting when the new passenger has very limited data.

7. Prediction for a new passenger

After adaptation, predictions for the new passenger are made using the posterior predictive.

Distribution. For a new input x, the prediction is demonstrated in equation (13):

$$\hat{y}_t(x) = \mathrm{E}_{q_t^*(\theta,\phi_t)}\left[f_{\theta,\phi_t}(x)\right].$$

The pseudocode for the proposed BNCD Algorithm is shown in Figure 1.

```
Algorithm: Transfer-Learning-Enabled BNCD for a New Passenger
Input:
   Source datasets D = {D_1, ..., D_N}
   Optional support set D_t^sup
   Optional passenger profile a_t
   Hyperparameters λ_reg, λ_KL, β, λ, τ, M
Output:
   Adapted posterior q_t^*
   Prediction function ŷ_t(x)
1. Preprocess source data:
     Normalize features
     Create passenger-level train/validation splits
2. Train source BNCD:
     Initialize θ, {ϕ_i}
     Minimize:
        L_S = Σ_i Σ_{(x,y)∈D_i} ℓ(f_{θ,ϕ_i}(x), y)
            + λ_reg R(θ, {ϕ_i})
            + λ_KL KL(q_S || p_0)
     Obtain θ_S, {ϕ_i^S}
     Store q_S and transfer statistics

3. Initialize new passenger t:
     If profile a_t is available then
        Compute weights:
           w_i = exp(s(a_t, a_i)/τ) / Σ_k exp(s(a_t, a_k)/τ)
        Set:
           ϕ_t0 = Σ_i w_i ϕ_i^S
           q_t0 = Σ_i w_i q_i^S
     Else
        Set:
           q_t0 = q_S
     End if
4. Adapt to new passenger:
     if D_t^sup is empty then
        q_t^* = q_t0
     Else
        Minimize:
           L_t = Σ_{(x,y)∈D_t^sup} ℓ(f_{θ,ϕ_t}(x), y)
               + β KL(q_t || q_t0)
               + λ || ϕ_t - ϕ_t0 ||^2
        q_t^* = argmin_{q_t} L_t
     End if
5. Predict:
     ŷ_t(x) = E_{q_t^*}[f_{θ,ϕ_t}(x)]
     Approximate by:
        ŷ_t(x) ≈ (1/M) Σ_{m=1}^M f_{θ^(m),ϕ_t^(m)}(x)
6. Online update:
     When new data arrive:
        q_t^new_prior = q_t^*
        Repeat adaptation step
        Optionally add passenger t to source pool
```

Figure 1. Pseudo code for the proposed BNCD algorithm.

IV. EXPERIMENTS

We examine the outcome of our optimized algorithm by conducting a dozen attempts in the field of overlapping communities. The results will be compared with existing methods, such as the SVD method [17] and the enhanced bipartite cluster method [17], to evaluate the performance of our optimized algorithm. The required environment to conduct the practical work for this paper is an Intel i5-core (TM) CPU, 8 GB of RAM, and a Windows 10 laptop. All iterations were conducted in the same environment and repeated 10 times; the average was then calculated. The proposed optimized Bipartite Algorithm is implemented in Python, and the dataset comprises three airline carrier datasets detailing every commercial flight in Libya, the

UAE, and Turkey. To evaluate the performance of the proposed work, various parameters were measured under different execution environments. We have proposed an Optimized Bipartite Algorithm. The proposed algorithm furnishes better performance in terms of Precision and Execution time. Comparative analysis in Table 2 and Figures 3 and 4 shows that the proposed algorithm is more efficient than the existing SVD algorithm and an improved bipartite clustering method.

V. CONTRIBUTIONS

This paper builds qualitative scientific contributions to the use of the optimized BNCD Algorithm, with a focus on the personalized recommendation experience as the passenger engages further with the platform:

1) Algorithmic Innovations: To improve the performance of recommendation techniques and figure out the cold start issue for new passengers and required data expandability issues.
2) Supporting passenger Experience :
   The proposed techniques efficiently personalize recommendations for each passenger using available historical data to enhance the passenger experience.

VI. DATASET AND EVALUATION METRICS

1. Used Dataset

The dataset comprises 3 carriers' datasets to examine the proposed approach: Emirates (EK), Afriqiyah Airways (8U), and Turkish Airlines (TK). Each dataset contains 1543 records of daily passenger reviews and ratings from January 1, 2016, to January 31, 2025, as shown in Figures 2, 3, and 4. The dataset used includes the following attributes:
Review titles: Emotions or specific keywords to be engaging.

- Author information: The information belongs to the passenger.
- dates: The published date of the review determined by the passenger.
- ratings: does the passenger give a score.
- status: is a credibility label used on a travel review website.
- Aircraft.: The type of aircraft operated.
- Travel Type: indicates the specific style,purpose , or category of a trip.
- Route: A flight from Source city to Destination city.
- Date Flown: (or flight date) refers to the day of the departed flight.
- Staff Service: They facilitate a smooth and comfortable journey for passengers.
- Food Quality: relies on (presentation, variety,and menu).
- Entertainment: comprises visual content, games, and the internet.
- WiFi: The modern Aircraft connects to satellites, providing coverage to passengers.
- Ground Services: An essential part of airport operations, and some tasks are performed while the plane is on the ground.
- Value for Money: balances ticket fare compared to the service quality.
- Recommended: Factor for rating the airline.
- Overall Rating:
- Review: a passenger’s evaluation of their flight experience.
- Country: A Reviewer’s Country.

SNO	Airline Name	Overall_F	Review_Title	Review Date	Verified	Review	Aircraft	Type Of T	Seat Typ	Route	Date Flown	Seat Com	Cabin Sta	Food & B	Ground S	Inflight E	W
8666	Emirates	1	"refusal to reimb	16th July 202	TRUE	The issu	A380-800	Family Le	Economy	New York t	Nov-22	1	1	1	1	1	
8667	Emirates	5	"Cabin crew were	16th July 202	TRUE	First the	A380	Solo Leis	Economy	Auckland t	May-23	4	1	1	4	5	
8668	Emirates	8	"letâ€™s see if th	6th July 2023	TRUE	What a	A380	Family Le	Business	Dubai to Sir	Jul-23	4	5	5	4	5	
8669	Emirates	5	"on-board intern	6th July 2023	TRUE	The on-board interne		Solo Leis	Economy	Dubai to Te	Jul-23	5	5	2	2	5	
8670	Emirates	1	"I was denied che	4th July 2023	TRUE	I'm writing this reviev		Solo Leis	Economy	Tokyo Narit	May-23	3	1	1	1	4	
8671	Emirates	6	"Emirates seems t	4th July 2023	TRUE	Very frie	A380	Family Le	Business	Munich to I	Jun-23	4	3	3	5	5	
8672	Emirates	1	"not even willing	29th June 20	TRUE	Did a single booking f		Family Le	Economy	Dubai to Lo	Jun-23	1	1	4	1	1	
8673	Emirates	7	"a hit or miss airli	29th June 20	TRUE	I decide	A380	Solo Leis	Economy	Dubai to Mi	Jun-23	3	3	3	2	5	
8674	Emirates	2	"Hope to see imp	26th June 20	TRUE	My fami	A380	Family Le	Economy	Dubai to Ho	Jun-23	3	2	1	3	3	
8675	Emirates	9	"appreciate the s	25th June 20	TRUE	This rev	Boeing 777	Solo Leis	Economy	Dubai to Co	Jun-23	4	5	4	3	4	
8676	Emirates	1	"no more Emirate	24th June 20	TRUE	I'm motivated to writ		Family Le	Economy	St Louis to I	Jun-23	1	1	2	1	2	
8677	Emirates	9	"a great experien	24th June 20	TRUE	Excellen	Boeing 777	Business	Economy	Birmingham	Jun-23	3	5	5	5	5	
8678	Emirates	4	"What a disappoi	17th June 20	TRUE	What a	A380	Business	Business	Glasgow to	May-23	4	2	2	3	4	
8679	Emirates	1	"sad, angry and h	13th June 20	FALSE	I didn't succeed to tal		Couple Le	Economy	Brussels to	Jun-23						
8680	Emirates	1	"certainly never f	10th June 20	TRUE	Emirates is an airline t		Solo Leis	Economy	Amman to	Jun-23	2	2	2	1	3	
8681	Emirates	1	"pleasant and prc	6th June 202	TRUE	Every e	A340	Couple Le	Economy	New York JF	Jun-23	5	5	5	5	5	
8682	Emirates	1	"excelling with th	6th June 202	TRUE	Birming	Boeing 777-3(	Solo Leis	Economy	Birmingham	Jun-23	4	5	5	4	5	
8683	Emirates	1	"worst experienc	31st May 202	TRUE	I had th	Boeing 777	Couple Le	Economy	Trivandrum	May-23	1	1	1	3	1	

Figure 3. Illustrates a sample of the historical daily Emirates Airways Reviews.

Sno	Airline Nar	Overall_	Review_Title	Review Da	Verifiec	Review	Aircraft	Type Of	Seat Ty	Route	Date Flow	Seat Co	Cabin St	Food &	Ground	Inflight	Wifi
21204	Turkish Air	4	"We had an ave	2nd May 2(	TRUE	We ha	A321	Busines	Econom	Nice to Ak	Apr-23	1	4	4	3	4	
21205	Turkish Air	5	"are not going t	2nd May 2(	TRUE	Worst communicat		Couple	Econom	Dallas For	Apr-23	2	4	2	1	3	
21206	Turkish Air	1	"assigned us all	1st May 20	FALSE	Very bad custome		Family I	Econom	New York	May-23	1	1	1	1	1	
21207	Turkish Air	5	"Not one smile	30th April 2	TRUE	Never I	A321	Couple	Econom	Istanbul t	Apr-23	1	4	3	1	5	
21208	Turkish Air	4	"The crew coul	29th April 2	TRUE	We fle	Boeing 787	Family I	Busines	Seattle to	Apr-23	3	2	2	2	1	
21209	Turkish Air	2	"donâ€™t fly th	28th April 2	TRUE	Origina	A333	Family I	Econom	Istanbul t	Apr-23	2	1	1	2	1	
21210	Turkish Air	1	"I waited for 6	24th April 2	TRUE	They don't allow e		Solo Lei	Econom	Budapest	Apr-23	1	1		1		
21211	Turkish Air	5	"internet on bc	20th April 2	TRUE	Ok flig	A330	Busines	Econom	Jeddah to	Apr-23	2	4	2	3		
21212	Turkish Air	1	"No leg room"	14th April 2	TRUE	Horrible experienc		Family I	Econom	San Franci	Apr-23	1	5	3	1	3	
21213	Turkish Air	1	"seats with no	10th April 2	TRUE	After booking and p		Couple	Econom	DÃ¼sseld	Apr-23	1	3	2	1		
21214	Turkish Air	1	"seats were so	9th April 2(	TRUE	The seats were so		Solo Lei	Econom	Dallas Ft V	Feb-23	1	1	1	2		
21215	Turkish Air	1	"extremely unl	8th April 2(	TRUE	I recently had an e>		Solo Lei	Econom	Delhi to D	Mar-23	1	1		1		
21216	Turkish Air	3	"I claimed to ge	8th April 2(	TRUE	I flew from Frankfu		Couple	Econom	Frankfurt	Apr-23	4	4	2	2	5	
21217	Turkish Air	1	"flight was ver	5th April 2(	FALSE	My flig	A320-200 /	Family I	Econom	Zurich to F	Jul-22	5	5	5	4	3	
21218	Turkish Air	1	"tell me my car	4th April 2(	TRUE	I checked in at Fra		Solo Lei	Econom	Frankfurt	Mar-23	4	5	4	1		

Figure 4. Illustrates a sample of the historical daily Turkish Airlines Reviews.

Sno	Airline Nar	Overall_	Review_Title	Review Da	Verifiec	Review	Aircraft	Type Of	Seat Ty	Route	Date Flow	Seat Co	Cabin St	Food &	Ground	Inflight	Wifi &
646	Afriqiyah A	8	Afriqiyah Airwa	23rd Augus	FALSE	Tripoli - London Gatwick - Tı		Economy Class									
647	Afriqiyah A	4	Afriqiyah Airwa	16th July 2(	FALSE	Dakar-Tripoli-Gatwick. Fantastic seats and legroom. Food was good as was in-flight entertainment. Frier											
648	Afriqiyah A	3	Afriqiyah Airwa	4th May 20	FALSE	Lagos to Tripoli was ok except the deplorable condition of the toilet no inflight entertainment cabin crev											
649	Afriqiyah A	3	Afriqiyah Airwa	27th April 2	FALSE	Dubai - Tripoli - Dakar. All in all the flight experience was a pleasure with very good food good seat spac											
650	Afriqiyah A	4	Afriqiyah Airwa	2nd Februa	FALSE	London - Tripoli - Accra. Have to say that in both directions a pleasant and trouble free flight. Onboard ca											
651	Afriqiyah A	3	Afriqiyah Airwa	6th January	FALSE	The airline has improved since I first started travelling with them4 years ago. The airplanes are new fooc											
652	Afriqiyah A	3	Afriqiyah Airwa	22nd Decer	FALSE	Gatwick to Accra. I was not disappointed with the quality of service - time seat cleanliness entertainmer											
653	Afriqiyah A	3	Afriqiyah Airwa	11th Nover	FALSE	Ouagadougou to Gatwick - the plane which was due to leave Ouaga at 1.00 am arrived at 4am & departe											
654	Afriqiyah A	2	Afriqiyah Airwa	7th Novem	FALSE	I travelled London Gatwick-Tripoli-Douala with Afriqiyah recently. Arriving 4 hours late at 01.00 hours I (											
655	Afriqiyah A	3	Afriqiyah Airwa	18th Septe	FALSE	Gatwick-Accra-Gatwick and found the journey up to expectations. On time both directions and the onbo											
656	Afriqiyah A	4	Afriqiyah Airwa	26th Augus	FALSE	London-Tripoli-Dakar returning in June. Much better than charter flights from Paris or charters to Gambi											
657	Afriqiyah A	2	Afriqiyah Airwa	17th June 2	FALSE	Afriqiyah Airways customer service is non-existent. I travelled from London via Tripoli to Cameroon last											
658	Afriqiyah A	1	Afriqiyah Airwa	30th May 2	FALSE	Have just flown to and from Tripoli. They were late leaving Gatwick on the outward journey due to missi											

Figure 5. Illustrates a sample of the historical Afriqiyah Airways Reviews.

2. Performance Evaluation Metrics

To assess the performance of the proposed approach, we use two key parameters: the Mean Absolute Error (MAE), given by equation (14), and Precision, given by equation (15), to measure the proximity of predicted grades to true ratings.

$$\text{Son} = \sum_{i-1}^{N} \frac{|r_{p,i} - \hat{r}_{p,i}|}{N}$$

$$\text{Precision} = \frac{\sum_{i=1}^{N} P_i}{N}$$

VII. RESULTS AND DISCUSSION

The λ parameter is chosen from the set {0.1, 0.5, 2, 4, 8, 12}, and the threshold value η of the network projection is set from the set { 0.3, 0.4, 0.5, 0.6, 0.7, 0.8}. The best way to gather groups that are in smaller patterns within the calculation network is the Special Community. 80 percent of the data is used for training the proposed approach, and 20 percent is used for testing. Table 2 demonstrates that the proposed Optimized Bipartite Algorithm consistently outperforms both SVD and the conventional Bipartite-cluster method across all three airline datasets (EK, TK, and 8U). It achieves the lowest values of 0.916, 0.964, and 0.782 for the EK, TK, and 8U datasets, respectively. On average, the proposed approach reduces the evaluation metric by 27.6% compared with SVD and 12.3% compared with the Bipartite-cluster method, indicating superior effectiveness and robustness across different datasets.

Table 2. Comparison of Mean Absolute Error (MAE) Results for SVD, Bipartite-Cluster, and the Proposed Optimized Bipartite Algorithm across Different Airline Datasets.

Dataset	**New Airline**		
	SVD	***Bipartite-cluster***	***Optimized Bipartite Algorithm***
EK	1.248	0.947	0.916
TK	1.319	1.291	0.964
8U	1.112	0.799	0.782

Table 3 compares the Mean Absolute Error (MAE) of SVD, Bipartite-cluster, and the proposed Optimized Bipartite Algorithm on the new passenger dataset. The results indicate that no single algorithm consistently achieves the best performance across all datasets. SVD produces the lowest MAE on the EK dataset (0.792), while the proposed Optimized Bipartite Algorithm performs best on the TK dataset with an MAE of 1.091. For the 8U dataset, the Bipartite-cluster algorithm achieves the lowest MAE (0.774). These findings suggest that the effectiveness of each algorithm depends on the characteristics of the dataset.

Table 3. Comparison of Mean Absolute Error (MAE) Results for SVD, Bipartite-Cluster, and the Proposed Optimized Bipartite Algorithm on the New Passenger Dataset.

Dataset	**New Passenger**		
	SVD	***Bipartite-cluster***	***Optimized Bipartite Algorithm***
EK	0.792	0.841	0.832
TK	1.273	1.339	1.091
8U	0.906	0.774	0.819

Table 4 compares the Mean Absolute Error (MAE) of SVD, Bipartite-cluster, and the proposed Optimized Bipartite Algorithm in the non-cold start scenario. The results show that the proposed algorithm achieves the lowest MAE for the EK (0.733) and TK (0.955) datasets, demonstrating superior prediction accuracy over

both SVD and the conventional Bipartite-cluster method. For the 8U dataset, however, the Bipartite-cluster algorithm slightly outperforms the proposed approach, recording the lowest MAE of 0.715 compared to 0.726 for the proposed algorithm and 0.731 for SVD. Overall, the proposed Optimized Bipartite Algorithm delivers the best performance on two of the three datasets, indicating that the optimization enhances recommendation accuracy under non-cold-start conditions while remaining competitive across datasets.

Table 4. Comparison of Mean Absolute Error (MAE) Results for SVD, Bipartite-Cluster, and the Proposed Optimized Bipartite Algorithm on the Non-Cold Start.

Dataset	**Non-Cold Start**		
	SVD	***Bipartite-cluster***	***Optimized Bipartite Algorithm***
EK	0.965	0.746	0.733
TK	1.218	1.122	0.955
8U	0.731	0.715	0.726

The proposed optimized Bipartite Algorithm yields better results than other algorithms. To verify the performance of the recommended airline, this paper also tests precision and runtime on the Emirates Airline dataset. Table 5 and Figure 6 show the accuracy of the three approaches in several suggested airlines. The number of recommended airlines is set to [5, 10, 15, 20]. It's clear that as the number of recommended airlines increases, the accuracy of the studied approaches decreases. However, the accuracy of the proposed optimized BNCD Algorithm remained relatively stable and slightly higher than that of other approaches.

Table 5. Precision Results.

Precision Result			
N	***Optimized Bipartite***	***SVD***	***Bipartite-Cluster***
5	0.38	0.34	0.3
10	0.36	0.33	0.29
15	0.35	0.32	0.28
20	0.34	0.31	0.27

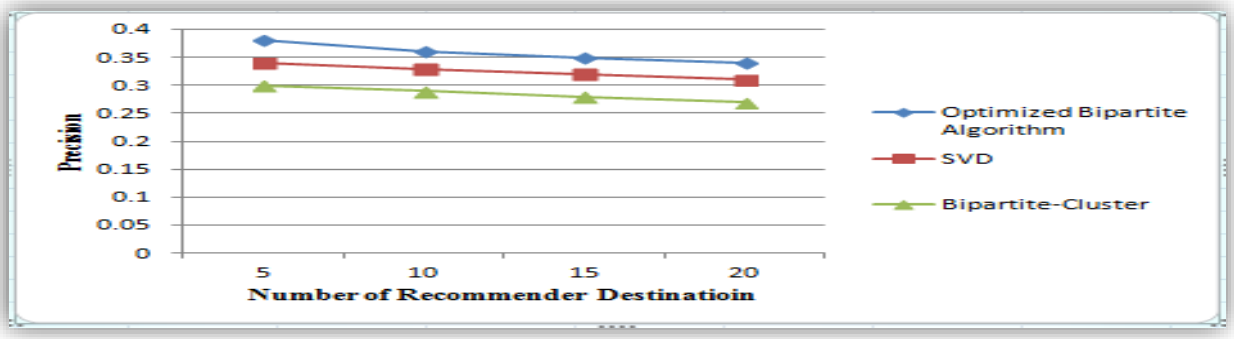


Figure 6. Precision Results

The Execution time of the Optimized BNCD Algorithm is demonstrated in Table 6 and Figure 7. To sum up, the optimized Bipartite Algorithm was noted to be capable when deployed to address the cold-start issue

in an Airline recommendation system. The optimized algorithm achieved a shorter runtime than other techniques.

Table 6. Execution Time (in seconds).

Execution time (in seconds)			
Iterations	***Optimized Bipartite***	***SVD***	***Bipartite-Cluster***
1	3	3	4
2	4	4	5
3	5	6	7
4	6	35	8
5	7	97	11
6	8	114	12

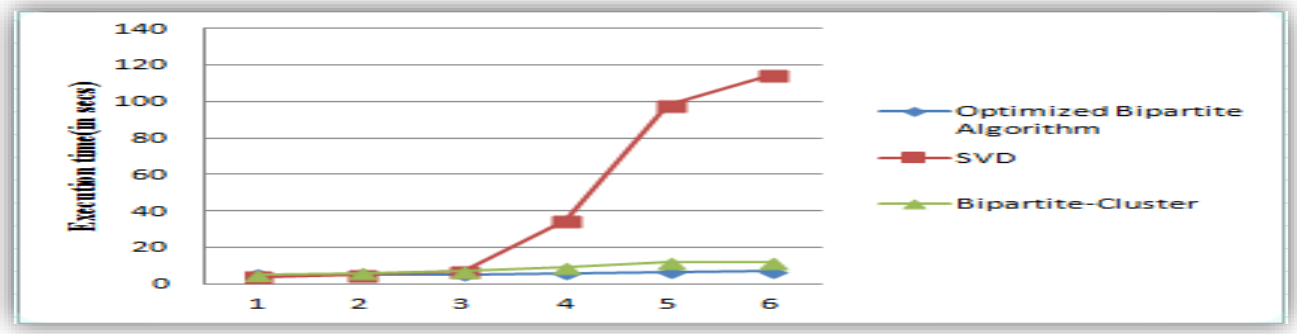


Figure 7. Performance Comparison and Analysis

VIII. CONCLUSION

In this paper, experimental results on three real airline evaluation datasets show that the proposed algorithm achieves higher recommendation accuracy and more stable performance. In future work, the algorithm will be further optimized to meet passenger needs, such as different destination classifications; it will also be tested on additional datasets to identify the structure in which passenger preferences overlap. We recommend implementing a comprehensive airline recommendation system and displaying the outcomes of preferred destination recommendations in a mobile App.

REFERENCES

[1]. Sandeep R. “Collaborative Filtering Techniques in Recommendation Systems[J]”. Data, Engineering and Applications, 2019, pp. 11-21.
[2]. Su X, Khoshgoftaar T M. “A Survey of Collaborative Filtering Techniques[J]”. Advances in Artificial Intelligence, 2009, 2009:1-19.
[3]. Ahn H J.” A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]”. Information Sciences, 2008, 178(1):37-51.
[4]. Koren Y, Bell R, Volinsky C. “Matrix Factorization Techniques for Recommender Systems”. Computer, 2009, 42(8):30-37.

[5]. Park, H.B, Y.S., Moon, S.C., Ko, H.H. “Clustering Method of Weighted Preference Using Means Algorithm and Bayesian Network for Recommender System[J]”. Journal of Information Technology Applications & Management Special Section, 2013. 219 – 230.
[6]. Fouss F, Pirotte A, Renders J M, et al. “Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation[J]”. IEEE Transactions on Knowledge and Data Engineering, 2007, 19: p.355-369.
[7]. Li X, Su X, Wang M. “Social network-based recommendation: a graph random walk kernel approach”. [C]// Acm/ieee-cs Joint Conference on Digital Libraries. ACM, 2012.
[8]. Seyerlehner K, Widmer G, Knees P. “A Comparison of Human, Automatic and Collaborative Music Genre Classification and User Centric Evaluation of Genre Classification Systems[C]”// Adaptive Multimedia Retrieval Context, Exploration, & Fusion International Workshop, Amr, Linz, Austria, August, Revised Selected Papers. Springer-Verlag, 2010.
[9]. Vall A, Eghbal-Zadeh H, Dorfer M, et al. “Music Playlist Continuation by Learning from Hand-Curated Examples and Song Features: Alleviating the Cold-Start Problem for Rare and Out-of-Set Songs[J]”. 2017.
[10]. Koren Y. “Factorization meets the neighborhood: A multifaceted collaborative filtering model”. In: Proc. of the 8th Int'l ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. ACM Press, 2008. 426-434.
[11]. Shi L, Zhao W X, Shen Y D. “Local Representative-Based Matrix Factorization for Cold-Start Recommendation[J]”. Acm Transactions on Information Systems, 2017, 36(2):1-28.
[12]. Ke-Han C. “User Clustering Based Social Network Recommendation[J]”. Chinese Journal of Computers, 2013.
[13]. Chao Z, Ying-Hua S, Hua-Feng X, et al. “Collaborative Filtering Recommendation Algorithm Based on Bidirectional Clustering on Users and Items[J]”. journal of qingdao university (natural science edition), 2018.
[14]. Xu B, Bu J, Chen C, et al.” An exploration of improving collaborative recommender systems via user-item subgroups[C]”// 2012.
[15]. Zhang Y, Zhang M, Liu Y, et al. “Improve collaborative filtering through bordered block diagonal form matrices[J]”. 2013.
[16]. Ratings P A, Jin R, Si L, et al. “Collaborative Filtering with Decoupled Models for[C]”// Twelfth International Conference on Information & Knowledge Management. ACM, 2003
[17]. Qin J, Ren K, Fang Y, et al. “Sequential Recommendation with Dual Side Neighbor-based Collaborative Relation Modeling[J]”.2019.
[18]. D. K. Panda and S. Ray, ‘‘Approaches and algorithms to mitigate cold start problems in recommender systems: A systematic literature review,’’ J. Intell. Inf. Syst., vol. 59, no. 2, pp. 341–366, Oct. 2022, doi: 10.1007/s10844-022-00698-5.
[19]. L. H. Son, ‘‘Dealing with the new user cold-start problem in recommender systems: A comparative review,’’ Inf. Syst., vol. 58, pp. 87–104, Jun. 2016, doi: 10.1016/j.is.2014.10.001.
[20]. N. A. Abdullah, R. A. Rasheed, M. H. N. M. Nasir, and M. M. Rahman, ‘‘Eliciting auxiliary information for cold start user recommendation: A survey,’’ Appl. Sci., vol. 11, no. 20, p. 9608, Oct. 2021, doi: 10.3390/app11209608.
[21]. H. Yuan and A. A. Hernandez, "User Cold Start Problem in Recommendation Systems: A Systematic Review," in IEEE Access, vol. 11, pp. 136958-136977, 2023, doi: 10.1109/ACCESS.2023.3338705.
[22]. Ke Yin Cao et al 2020"Improving the cold start problem in Music Recommender System ", J. Phys.: Conf. Ser. 1651 012067.
[23]. Z. Zhang, M. Dong, K. Ota and Y. Kudo, "Alleviating New User Cold-Start in User-Based Collaborative Filtering via Bipartite Network," in *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 672-685, June 2020, doi: 10.1109/TCSS.2020.2971942.

[24]. A. N. S, G. R. G, N. R. N, and H. P. K, "Tourism Recommendation System Using Hybrid Approach," *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2025, pp. 1039-1043, doi: 10.1109/ICIRCA65293.2025.11089855.

[25]. Karthiyayini, J. "Personalized Travel Recommendations System Using Hybrid Filtering and Deep Learning." *Journal of Information Systems Engineering and Management*, 2025, https://doi.org/10.52783/jisem.v10i13s.2011.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

The Impact of Syntactic Dependency Information on Turkish Coreference Resolution

Ebru Çıracı
Department of Computer Engineering
Faculty of Engineering, Architecture and Design,
Bartın University
Bartın, Türkiye
esomuncu@bartin.edu.tr
ORCID:0000-0003-0730-0682

Nesrin Aydın Atasoy
Department of Computer Engineering
Faculty of Computer and Information Sciences,
Karabük University
Karabük, Türkiye
nesrinaydin@karabuk.edu.tr
ORCID:0000-0002-7188-0020

Abstract

Coreference resolution aims to identify and link mentions of the same entity in a text. While contextual language models produce strong representations for this task, they do not explicitly model the syntactic relations between words. In this study, the contribution of relation-labeled dependency information to Turkish coreference resolution is investigated. To this end, a relation-aware Graph Attention Network model was developed that integrates contextual word representations with syntactic representations derived from a dependency graph. The effects of the number of attention heads and enhanced dependency links for empty nodes on performance and stability were compared. All syntax-aware configurations outperformed the baseline model; the best configuration achieved an absolute improvement of 1.69 points with an average CoNLL F1 of 66.71%. The results indicate that relation-labeled dependency graphs support Turkish coreference resolution; however, increasing the number of attention heads and the number of empty node links does not yield a consistent additional gain.

Keywords: Turkish coreference resolution, relation-aware graph attention network, dependency graph, syntactic information, contextual representations

I. INTRODUCTION

Coreference resolution links information scattered throughout a document by grouping mentions of the same real-world entity or event into a single chain. Therefore, coreference resolution is considered an important component that supports various natural language processing tasks, such as question answering [1], text summarization [2], [3], and information extraction [4]. End-to-end neural models and pre-trained language encoders have significantly improved coreference resolution performance. While these models can learn context-aware representations of words, they often do not directly use explicit syntactic relations between words as a structured source of information. Consequently, representations obtained from the linear context of the text may not always explicitly reflect the dependency relations within a sentence. This limitation becomes more pronounced in Turkish, which has free word order and a rich morphological structure. Subject and possessive elements that are not explicitly present on the surface (dropped pronouns) can participate in coreference chains. Turkish-ITCC represents these dropped mentions using empty nodes and the relations between words through dependency annotations [5]. This structure allows modeling both syntactic connections between explicit mentions and relations between mentions not present on the surface and other nodes. However, the contribution of dependency relation labels and enhanced dependency links for empty nodes to Turkish coreference resolution has not been sufficiently investigated. In this study, a Graph Attention Network approach is examined that integrates contextual word representations from BERTurk with syntactic representations from

a relation-labeled dependency graph. Words and empty nodes are modeled as graph nodes, and dependency relations are modeled as labeled directed edges. The baseline model and graph configurations with different numbers of attention heads are compared across three random seeds; furthermore, the impact of DEPS links associated with empty nodes on performance and stability is evaluated.

II. RELATED WORK

Early learning-based studies on coreference resolution relied heavily on handcrafted shallow syntactic and semantic features. These approaches were developed as mention-pair models [6], [7], [8], that classify the coreference relation between pairs of mentions, entity-mention models [9], [10] that associate mentions with existing entity sets, and methods [11], [12] that rank candidate antecedents according to their relative suitability. Subsequent studies utilized neural representation learning methods in these processes to reduce reliance on hand-designed features [13], [14], [15], [16]. Following these developments, end-to-end neural models that learn mention detection and coreference linking within a joint architecture have come to the fore. Lee et al. [17] proposed an end-to-end model that treats possible text spans in a document as mention candidates and scores candidate antecedents for each mention. This approach was extended to support higher-order inference by reincorporating information from predicted coreference links into the mention representations [18]. Joshi et al. [19] showed that the SpanBERT model, which focuses on representing text spans, improves coreference resolution performance. However, most of these studies have focused on resource-rich languages; the detection of dropped mentions, which are not explicitly present on the surface in Turkish, and their linking to coreference chains have been examined to a more limited extent. A representation and evaluation approach for including dropped subject and possessive pronouns in coreference resolution has been presented for Turkish; in the Turkish-ITCC dataset, these mentions are represented by empty nodes and marked within the same coreference chains as explicit mentions [5], [20]. Thus, the dataset provided an evaluation environment that requires identifying mentions not present in the surface text and adding them to appropriate coreference chains, in addition to linking explicit mentions. Turkish-ITCC is part of CorefUD, a multilingual coreference collection compatible with Universal Dependencies [21]. Systems developed in different languages and on different datasets were compared in shared tasks organized around this collection [22]. In the CRAC 2024 shared task using CorefUD 1.2, empty nodes were removed from the system input, and models were expected to predict these nodes in addition to performing coreference resolution [23]. Straka [24] examined single- and two-stage CorPipe configurations that include empty-node prediction in the coreference resolution process. Pražák and Konopík [25], on the other hand, developed an end-to-end multilingual model that represents mentions by their head words and discussed the possible contribution of syntactic information to this representation. Although pre-trained language models produce strong contextual representations, explicit dependency links in datasets are not always transferred to these representations as structured information. Dependency parsing provides a natural graph structure in which words are represented as nodes and syntactic relations as edges. Graph Attention Networks weight the information transferred from neighboring nodes with learnable attention coefficients, ensuring the selective transfer of graph connections to node representations [26]. Previous studies have demonstrated the effectiveness of graph-based structures for coreference resolution. Jiang and Cohn [27] integrated dependency-based syntactic information with semantic role relations in a heterogeneous graph attention network, improving performance on the English and Chinese OntoNotes 5.0 datasets. Similarly, Yousfi et al. [28] proposed a heterogeneous graph-based model that combines dependency and semantic role information for the Arabic CoNLL-2012 dataset, outperforming BERT-based methods and achieving competitive results compared with sequence-to-sequence approaches. While standard graph attention networks learn the relative importance of neighboring nodes, they do not always model different edge types separately. Relation-aware graph attention networks address this by incorporating edge-type information into the attention and information-transfer processes. Meng et al. [29] combined BERT-based contextual representations with dependency-based syntactic representations using a relation-aware graph attention network. Experiments on



GAP and OntoNotes 5.0 showed that preserving dependency relation types can support coreference resolution performance.

In the CoNLL-U format, the HEAD and DEPREL fields represent the basic dependency tree, while the DEPS field represents enhanced dependency links added to this tree. Enhanced dependency representations have been reported to more explicitly encode syntactic relations that are not directly present in the basic tree [30], [31]. In Turkish-ITCC, the DEPS field is important because it can include syntactic connections between empty nodes, which have no explicit surface-word counterpart, and other nodes. However, the effect of transferring these links to the coreference model on performance remains unclear. Previous studies show that dependency information can be transferred to coreference resolution using graph-based methods and that modeling relation labels can support contextual representations. However, the contribution of enhanced dependency links (DEPS) associated with empty nodes in Turkish-ITCC, and the performance and stability of the dependency graph across different numbers of attention heads, have not been sufficiently compared. This study evaluates the contributions of the relation-labeled dependency graph, the effect of empty node links, and the model behavior across different attention head configurations.

III. MODEL

In this study, an end-to-end Turkish coreference resolution architecture is proposed that integrates contextual word representations with relation-labeled dependency information. The model consists of contextual encoding, syntactic graph processing, representation integration, mention detection, and coreference linking stages. The overall architecture is shown in Figure 1.

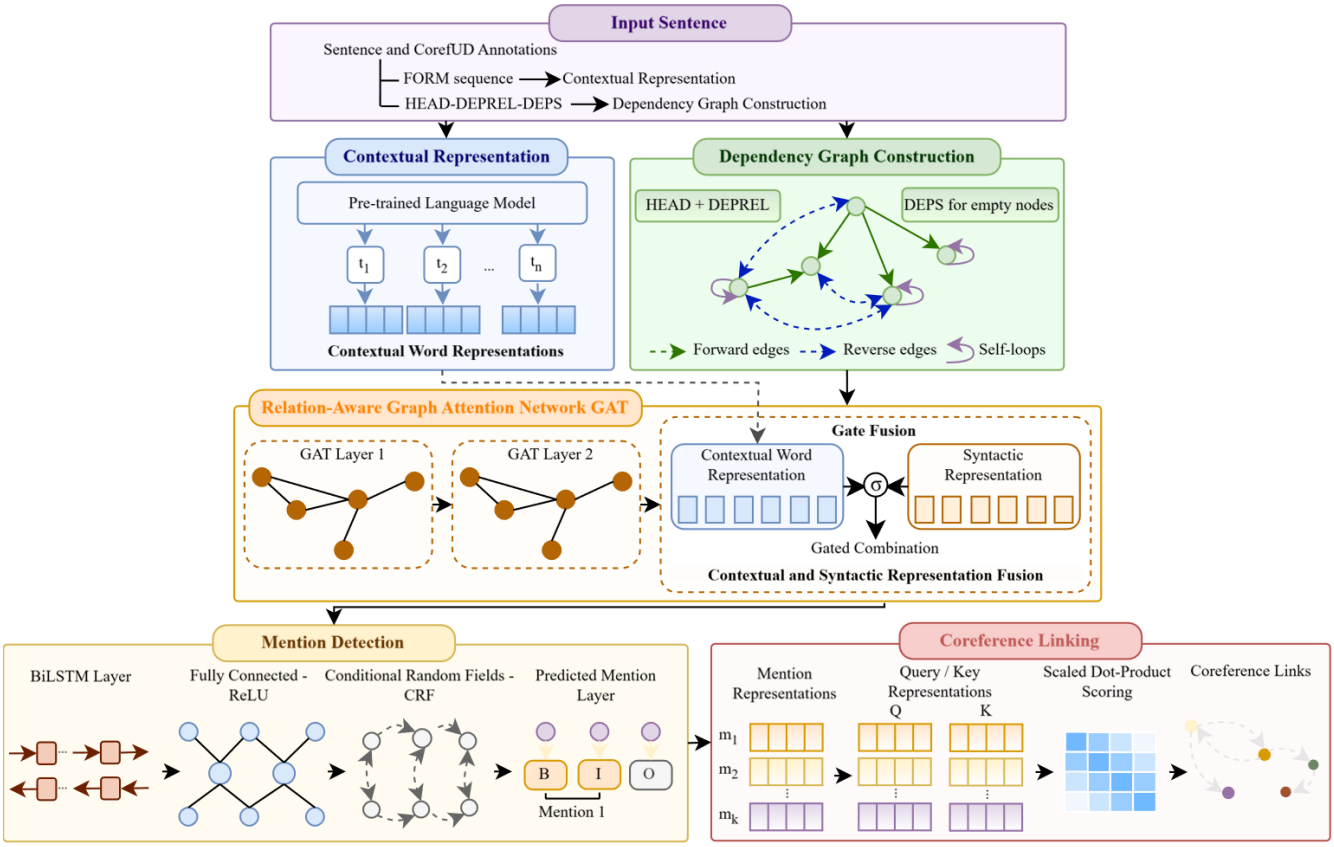


Figure 1. Overall architecture of the proposed relation-aware dependency graph attention model.

The input text is first transformed into contextual word representations ($i.$) using the BERTurk [32] encoder. In parallel, the words in the sentence and ttheir dependency relations are represented as a relation-labeled directed graph (). This graph structure is processed through a two-layer Relation-Aware Graph Attention Network (RGAT) to produce representations (i) eflecting the syntactic roles of the words:

$$s_i = RGAT_2(RGAT_1(x_i, G), G) \tag{1}$$

Contextual information ($i.$) $and\ syntactic\ information$ ($s\ alt\ i.$) obtained from the dependency graph are combined through a learnable gating mechanism. This mechanism dynamically balances the impact of both information sources on the final representation. The integrated node representation (f_i) is obtained as follows:

$$f_i = g_i \odot x_i + (1 - g_i) \odot s_i \tag{2}$$

Here, g_i denotes the gate value that determines the relative contribution of both sources and is calculated as follows:

$$g_i = \sigma\left(W_g[x_i; s_i] + b_g\right) \tag{3}$$

In this formula, σ represents the sigmoid activation function, and W_g and b_g represent learnable parameters.

Integrated representations are fed into a CRF [33] layer operating over two BiLSTM layers and BIO labels to determine mention boundaries. Identified mentions (m_i) are encoded by concatenating the representations at the start and end positions. In the final stage, the coreference similarity score (s_{ij}) between each i-th mention and candidate antecedents (j) is calculated using a query-key-based mechanism:

$$q_i = W_q m_i + b_q, \quad k_j = W_k m_j + b_k \tag{4}$$

$$s_{ij} = \frac{q_i^T k_j}{\sqrt{d_k}} \tag{5}$$

The model links each mention to the most appropriate coreference chain or starts a new chain based on these calculated scores.

IV. EXPERIMENTAL SETUP

1. Dataset and Evaluation Protocol

Experiments were conducted on the Turkish-ITCC dataset in the CorefUD 1.2 collection. Excluding single coreference sets, the dataset contains 24 documents, 4,732 sentences, 55,358 non-empty words, 11,584 empty nodes, 4,019 coreference sets, and 21,569 mentions [24]. The published data split consists of 19 training, two development, and three test documents [20]. Since gold test annotations are not publicly available, comparisons were conducted on the development set; performance was evaluated with MUC [34], B^3 [35], CEAFe [36], and the average of these three metrics, CoNLL F1 [37]. Each model was trained with three different random seed values (42, 43, and 44), and results are reported as mean ± standard deviation.

2. Training Settings and Model Configurations

Models were trained using the dbmdz/bert-base-turkish-128k-uncased encoder for 15 epochs, with a batch size of 8, a sequence length of 512, a learning rate of 2×10^{-5}, a dropout rate of 0.2, and LazyAdam optimization. In Syntax-GAT configurations, two relation-aware GAT layers and 64-dimensional relation embeddings were used. All experiments were conducted using TensorFlow on an NVIDIA GeForce RTX 5060 Ti GPU with 16 GB of memory. The differences in configurations, including empty node DEPS links and the number of attention heads, are presented in Table 1. Computational cost was assessed using mean runtime and peak GPU memory consumption. Mean runtime was calculated across seeds 42–44.

Table 1. Evaluated model configuration

Model	***Empty-node DEPS***	***GAT layers***	***Attention heads***
Baseline	N/A	-	-
Syntax-GAT-DEPS-1H	Yes	2	1
Syntax-GAT-DEPS-4H	Yes	2	4
Syntax-GAT-1H	No	2	1
Syntax-GAT-4H	No	2	4

V. RESULTS AND DISCUSSION

1. Overall Performance

The average MUC, B^3, CEAFe, and CoNLL F1 scores obtained across three random seeds for the model configurations are presented in Table 2. Results are reported as mean ± standard deviation.

Table 2. Mean coreference performance ± standard deviation across three random seeds

Model	*MUC F1 (%)*	*B^3 F1 (%)*	*CEAFe F1 (%)*	*CoNLL F1 (%)*	*Gain over Baseline*
Baseline	81.42 ± 0.19	54.58 ± 1.65	59.06 ± 0.95	65.02 ± 0.66	-
Syntax-GAT-DEPS-1H	82.02 ± 0.19	56.40 ± 2.56	59.08 ± 0.60	65.84 ± 1.02 [a]	+0.82
Syntax-GAT-DEPS-4H	82.10 ± 0.20	55.39 ± 2.32	59.94 ± 1.10	65.80 ± 1.12	+0.78
Syntax-GAT-1H	82.26 ± 0.20	57.80 ± 3.78	60.06 ± 0.49	66.71 ± 1.41 [a]	+1.69
Syntax-GAT-4H	81.84 ± 0.65	56.35 ± 0.35	59.66 ± 1.00	65.95 ± 0.64	+0.93

[a.] Statistically significant improvement over the baseline according to a paired one-tailed t-test ($p < 0.05$).

Experimental results reveal that all proposed Syntax-GAT configurations outperform the baseline model. Models supported with syntactic information achieved absolute increases of 0.78-1.69 points in the CoNLL F1 score compared to the baseline model. The Syntax-GAT-1H configuration, which achieved the highest average performance, reached an average CoNLL F1 score of 66.71%. These findings suggest that representations obtained from relation-labeled dependency graphs significantly complement contextual word representations, thereby increasing Turkish coreference resolution performance. The paired one-tailed t-test indicated that the improvements achieved by the Syntax-GAT-DEPS-1H ($p = 0.049$) and the best-performing Syntax-GAT-1H ($p = 0.036$) configurations over the baseline model were statistically significant.

2. Effect of Attention Heads and Empty-Node DEPS

CoNLL F1 results based on seeds are given in Table 3. In models not containing DEPS, the single-headed configuration performed 0.76 points higher than the four-headed configuration, while the difference between models containing DEPS is only 0.04 points.

Table 3. CoNLL F1 results for different random seeds.

Model	*Seed 42*	*Seed 43*	*Seed 44*	*Mean CoNLL F1 (%) ± SD*	*Mean runtime (min)*	*Peak GPU memory (GB)*
Baseline	65.32	65.48	64.26	65.02 ± 0.66	91.55	8.88
Syntax-GAT-DEPS-1H	66.68	66.12	64.71	65.84 ± 1.02	105.51	9.18
Syntax-GAT-DEPS-4H	65.66	66.99	64.76	65.80 ± 1.12	104.08	9.00
Syntax-GAT-1H	67.85	67.14	65.13	66.71 ± 1.41	109.66	9.26
Syntax-GAT-4H	65.85	65.37	66.63	65.95 ± 0.64	108.02	9.17

The computational results indicate that the Syntax-GAT configurations introduce a limited additional cost compared with the baseline model. When models with the same number of attention heads are compared, configurations without DEPS links exhibit longer mean training times than their DEPS-based counterparts. Relative to the baseline, the mean training time increased by approximately 13.7–19.8%, while peak GPU

memory usage increased by only 1.4–4.3%. Increasing the number of attention heads from one to four did not consistently improve average performance, suggesting that distributing information from the same dependency neighborhood across multiple heads may not produce sufficiently complementary patterns. However, Syntax-GAT-4H was the most stable graph-based configuration with a standard deviation of 0.64, indicating that multiple heads may contribute more to reducing seed-related variability than to improving average performance. Similarly, adding DEPS links for empty nodes did not yield further gains; CoNLL F1 decreased by 0.87 points in the single-head model and by 0.15 points in the four-head model compared with their counterparts without DEPS. This suggests that DEPS links may introduce information mixing by transferring less informative connections along with useful syntactic relations into the node representations.

3. Stability and Discussion

When model behavior across different random seed values is examined, the Syntax-GAT-1H configuration, which achieves the highest average performance, also has the highest standard deviation (1.41). Although Syntax-GAT-4H stands out as the most stable configuration, the highest performance and the highest stability did not converge in the same model. The fact that results are based on a development set comprising only two documents and three different seed values constitutes a fundamental limitation on the generalizability of the findings.

VI. CONCLUSION

This study aims to investigate the effect of relation-labeled dependency information on Turkish coreference resolution by integrating contextual representations with a Graph Attention Network architecture. The proposed architecture outperformed the baseline model across all scenarios by explicitly modeling syntactic roles. The most significant quantitative result is that the best configuration achieved an absolute increase of 1.69 points in CoNLL F1 over the baseline model. In general, the structured transfer of dependency relations improves performance, but increasing the number of attention heads and the number of empty node links does not provide additional benefit. The main limitation of the study is that the evaluation was performed on a limited development set. Future work will evaluate the proposed approach on additional datasets and other morphologically rich languages to further examine its generalizability. A more detailed error analysis will also investigate which types of coreference cases benefit most from dependency-based syntactic information and in which cases the model continues to fail.

REFERENCES

[1]. S. Bhattacharjee, R. Haque, G. M. de Buy Wenniger, and A. Way, “Investigating query expansion and coreference resolution in question answering on BERT,” in *Natural Language Processing and Information Systems*, vol. 12089, E. Métais, F. Meziane, H. Horacek, and P. Cimiano, Eds. Cham, Switzerland: Springer, 2020, pp. 47–59, doi: 10.1007/978-3-030-51310-8_5.

[2]. Z. Liu, K. Shi, and N. Chen, “Coreference-aware dialogue summarization,” in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore and Online: Association for Computational Linguistics, 2021, pp. 509–519, doi: 10.18653/v1/2021.sigdial-1.53.

[3]. J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Ježek, “Two uses of anaphora resolution in summarization,” *Information Processing & Management*, vol. 43, no. 6, pp. 1663–1680, Nov. 2007, doi: 10.1016/j.ipm.2007.01.010.

[4]. S. Kriman and H. Ji, “Joint detection and coreference resolution of entities and events with document-level context aggregation,” in *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, Online: Association for Computational Linguistics, 2021, pp. 174–179, doi: 10.18653/v1/2021.acl-srw.18.

[5]. T. Pamay Arslan and G. Eryiğit, "Enhancing Turkish coreference resolution: Insights from deep learning, dropped pronouns, and multilingual transfer learning," *Computer Speech & Language*, vol. 89, Art. no. 101681, 2025, doi: 10.1016/j.csl.2024.101681.

[6]. V. Ng, "Machine learning for coreference resolution: From local classification to global ranking," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI, USA: Association for Computational Linguistics, 2005, pp. 157–164, doi: 10.3115/1219840.1219860.

[7]. P. Denis and J. Baldridge, "Joint determination of anaphoricity and coreference resolution using integer programming," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, NY, USA: Association for Computational Linguistics, 2007, pp. 236–243.

[8]. I. Haponchyk and A. Moschitti, "A practical perspective on latent structured prediction for coreference resolution," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain: Association for Computational Linguistics, 2017, pp. 143–149, doi: 10.18653/v1/E17-2023.

[9]. A. McCallum and B. Wellner, "Conditional models of identity uncertainty with application to noun coreference," in *Advances in Neural Information Processing Systems 17*, 2004, pp. 905–912.

[10]. A. Culotta, M. Wick, and A. McCallum, "First-order probabilistic models for coreference resolution," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, NY, USA: Association for Computational Linguistics, 2007, pp. 81–88.

[11]. P. Denis and J. Baldridge, "A ranking approach to pronoun resolution," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, Hyderabad, India, 2007, pp. 1588–1593.

[12]. A. Rahman and V. Ng, "Ensemble-based coreference resolution," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Volume Three*, Barcelona, Spain: AAAI Press, 2011, pp. 1884–1889.

[13]. S. Martschat and M. Strube, "Latent structures for coreference resolution," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 405–418, 2015, doi: 10.1162/tacl_a_00147.

[14]. K. Clark and C. D. Manning, "Entity-centric coreference resolution with model stacking," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, 2015, pp. 1405–1415, doi: 10.3115/v1/P15-1136.

[15]. S. Wiseman, A. M. Rush, S. Shieber, and J. Weston, "Learning anaphoricity and antecedent ranking features for coreference resolution," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, 2015, pp. 1416–1426, doi: 10.3115/v1/P15-1137.

[16]. S. Wiseman, A. M. Rush, and S. M. Shieber, "Learning global features for coreference resolution," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA: Association for Computational Linguistics, 2016, pp. 994–1004, doi: 10.18653/v1/N16-1114.

[17]. K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 188–197, doi: 10.18653/v1/D17-1018.

[18]. K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2, Short Papers*, New Orleans, LA, USA: Association for Computational Linguistics, 2018, pp. 687–692, doi: 10.18653/v1/N18-2108.

[19]. M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020, doi: 10.1162/tacl_a_00300.

[20]. T. Pamay Arslan and G. Eryiğit, "Incorporating dropped pronouns into coreference resolution: The case for Turkish," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 14–25, doi: 10.18653/v1/2023.eacl-srw.2.

[21]. A. Nedoluzhko, M. Novák, M. Popel, Z. Žabokrtský, A. Zeldes, and D. Zeman, "CorefUD 1.0: Coreference meets Universal Dependencies," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 2022, pp. 4859–4872.

[22]. Z. Žabokrtský *et al*., "Findings of the second shared task on multilingual coreference resolution," in *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution, Singapore: Association for Computational Linguistics*, 2023, pp. 1–18, doi: 10.18653/v1/2023.crac-sharedtask.1.

[23]. M. Novák *et al*., "Findings of the third shared task on multilingual coreference resolution," in *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, Miami, FL, USA: Association for Computational Linguistics, 2024, pp. 78–96, doi: 10.18653/v1/2024.crac-1.8.

[24]. M. Straka, "CorPipe at CRAC 2024: Predicting zero mentions from raw text," in *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, Miami, FL, USA: Association for Computational Linguistics, 2024, pp. 97–106, doi: 10.18653/v1/2024.crac-1.9.

[25]. O. Pražák and M. Konopík, "End-to-end multilingual coreference resolution with headword mention representation," in *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, Miami, FL, USA: Association for Computational Linguistics, 2024, pp. 107–113, doi: 10.18653/v1/2024.crac-1.10.

[26]. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada, 2018.

[27]. F. Jiang and T. Cohn, "Incorporating syntax and semantics in coreference resolution with heterogeneous graph attention network," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, 2021, pp. 1584–1591, doi: 10.18653/v1/2021.naacl-main.125.

[28]. R. E. Yousfi, M. Ourekouch, and M. Erradi, "Heterogeneous graph neural network based Arabic coreference resolution," in *Networked Systems: 13th International Conference, NETYS 2025*, Rabat, Morocco: Springer, 2026, pp. 253–270, doi: 10.1007/978-3-032-00347-8_17.

[29]. Y. Meng, X. Pan, J. Chang, and Y. Wang, "RGAT: A deeper look into syntactic dependency information for coreference resolution," in *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*, Gold Coast, Australia: IEEE, 2023, pp. 1–8, doi: 10.1109/IJCNN54540.2023.10191577.

[30]. S. Schuster and C. D. Manning, "Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia: European Language Resources Association, 2016, pp. 2371–2378.

[31]. J. Nivre *et al*., "Enhancing Universal Dependency treebanks: A case study," in *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 102–107, doi: 10.18653/v1/W18-6012.

[32]. S. Schweter, "BERTurk-BERT models for Turkish," Zenodo, 2020, doi: 10.5281/zenodo.3770924.



[33]. J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, MA, USA, 2001, pp. 282–289.

[34]. M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, "A model-theoretic coreference scoring scheme," in *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, USA, 1995, pp. 45–52, doi: 10.3115/1072399.1072405.

[35]. A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in *Proceedings of the LREC 1998 Workshop on Linguistic Coreference*, Granada, Spain, 1998, pp. 563–566.

[36]. X. Luo, "On coreference resolution performance metrics," in *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, BC, Canada: Association for Computational Linguistics, 2005, pp. 25–32.

[37]. S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," in *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*, Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 1–40.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Physics-Informed Hybrid Framework for Predictive Supervision and Multi-Objective Optimization of Ultrafiltration Membrane Fouling in Desalination

Zineb Idabella
Cadi Ayyad University (UCA), Faculty of Sciences and Techniques, Laboratory of Computer Engineering and Systems (L2IS), BP 549, Avenue Abdelkrim El Khattabi, Marrakesh 40000, Morocco
z.idabella.ced@uca.ac.ma

Laila Amir
Cadi Ayyad University (UCA), Faculty of Sciences and Techniques, Laboratory of Computer Engineering and Systems (L2IS), BP 549, Avenue Abdelkrim El Khattabi, Marrakesh 40000, Morocco
l.amir@uca.ac.ma

Abstract

Membrane fouling remains a major challenge in ultrafiltration (UF) pre-treatment for seawater reverse osmosis (RO) desalination, leading to increased energy consumption, excessive chemical usage, and accelerated membrane deterioration. This study presents a physics-informed predictive supervision framework operating at the level of individual membrane modules, in contrast to conventional reactive strategies based on rack-average monitoring. The proposed framework combines a dynamic physicochemical resistance model, numerically solved using a fourth-order Runge–Kutta (RK4) scheme, with a multi-output Temporal Convolutional Network (TCN) trained on a hybrid dataset integrating real industrial measurements and simulation-generated features. The framework was evaluated on a publicly available 150-hour operational benchmark using a chronological train/test split, achieving predictive performance of up to $R^2 = 0.97$ for coagulant dosing. Comparative simulation results indicate potential reductions of up to 36.6% in maintenance interventions, 75.99% in chemical consumption, and 32.16% in energy demand, together with an estimated 60.70% increase in membrane service life. As these findings are based on a simulation-enriched study conducted at a single site, they should be interpreted as indicative upper-bound estimates pending validation under broader experimental conditions.

Keywords: Digital twin; Physics-informed deep learning; Membrane fouling; Dynamic coagulant dosing; Multi-objective optimization

I. INTRODUCTION

Global freshwater scarcity has made seawater desalination, and reverse osmosis (RO) in particular, a cornerstone of water security [1]. The long-term performance of RO plants depends critically on ultrafiltration (UF) pre-treatment, which acts as the primary barrier against particulate matter, colloids, and biological contaminants. As filtration progresses, foulant accumulation induces cake layer formation and pore constriction, hydraulically manifested as an exponential rise in transmembrane pressure (TMP) and specific resistance. Counteracting this degradation requires intensive backwashing and coagulant-based pre-treatment [2], [3], which introduces cascading inefficiencies: mechanical fatigue on membrane fibers and escalating operating expenditures (OPEX) from overdosing. Deterministic resistance-in-series models provide a physically interpretable baseline [4], [5] but struggle with the nonlinear, time-varying nature of fouling [6], [7]. AI-based approaches have substantially expanded the predictive toolkit [8], [9], with recent deep learning and reinforcement learning methods better capturing temporal dynamics [10], [11]. A critical limitation, however, persists: the tacit assumption that parallel UF modules foul uniformly. Existing adaptive strategies [12], [13] compute resistance and trigger backwash

events based on rack-wide averages, thereby masking the heterogeneous fouling trajectories arising from hydrodynamic imbalances. The present work addresses this gap by employing a hybrid physics-informed TCN framework that monitors each module individually, enabling proactive, asynchronous supervisory control.

Recent work reinforces this rack-level bias even as predictive tools have grown more sophisticated. [14] trace the shift from empirical fouling indices to intelligent, data-driven predictors, while [15] embed fouling mechanisms directly into machine-learning architectures to improve interpretability and generalization. [16] demonstrated a real-time deep-learning framework for hybrid NF/RO plants, predicting flux, salt rejection, and energy consumption with R^2 above 0.95, and noted its potential as the computational backbone of full-scale digital twins. However, like the self-adaptive rack-level controllers of [12], [13], these frameworks still reason over aggregate or plant-wide signals: none resolve the module-to-module heterogeneity documented in Figure 1, nor translate individual-module forecasts into an actionable, asynchronous control policy. This is precisely the gap the present framework is designed to close.

II. METHODOLOGY

1. Data Description and Window Selection

Operational data were sourced from the publicly available Dryad repository [17], comprising 422 days of UF system operation as pre-treatment for a seawater RO plant. The dataset records feed pressure, TMP, individual module flow rates, and feed water temperature at high temporal resolution. A contiguous 150-hour window was deliberately selected as the modeling baseline. The original 422-day record contains intermittent logging gaps and fragmented daily sequences; both the RK4 state-space integration and the TCN require unbroken chronological data to preserve the physical integrity of filtration/backwash cycles. The selected 150-hour window is the longest uninterrupted sequence, spanning multiple fouling phases and backwash events that are essential for reconstructing transient behavior. The real sensor signals for individual module throughput are displayed in Figure 1, showing that structural and manifold imbalances generate measurable, module-specific flow deviations even under nominally identical operating conditions—a physical reality that rack-average controllers cannot resolve.

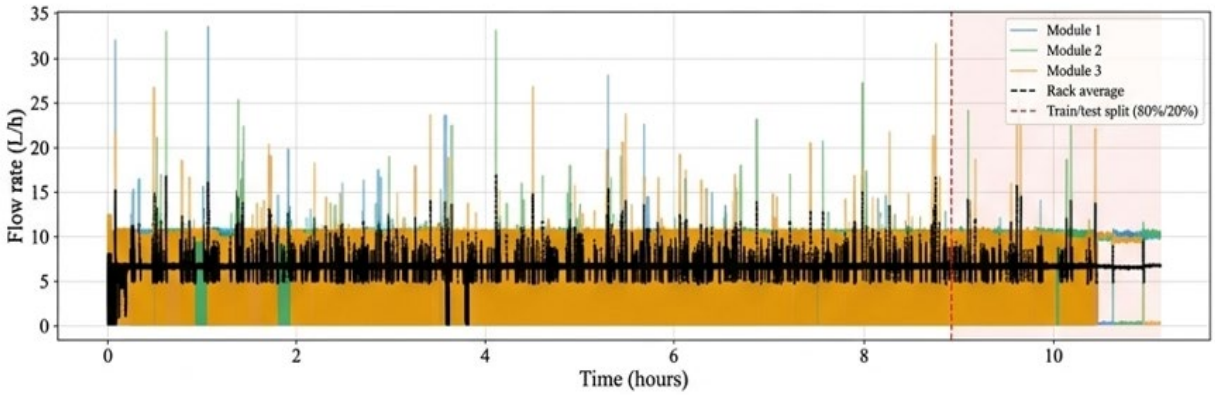


Figure 1. Variation in influent flow rates across identical parallel UF modules, illustrating the hydrodynamic heterogeneity that leads to module-specific fouling trajectories

2. Module Level Mathematical Simulator and Parameter Calibration

A dynamic mathematical simulator was developed to track the individual hydraulic resistance Ri(t) of each parallel module, distinguishing reversible cake-layer resistance Rcake, i, from irreversible fouling resistance Rirr, i, and incorporating a real-time coagulant dosing strategy C(t):

$$\begin{cases} dR_{cake,i}/dt = \beta_{cake} \cdot \eta_i \cdot J_i(t)\cdot(1 + 0.04\cdot C(t)) \\ dR_{irr,i}/dt = \beta_{irr} \cdot \eta_i \cdot J_i(t)\cdot\exp(-0.05\cdot C(t)) \\ R_i(t) = R_m + R_{cake,i}(t) + R_{irr,i}(t) \\ J_i(t) = \Delta P_i(t) / [\mu(t)\cdot R_m^{ref}\cdot R_i(t)] \cdot \delta_i(t) \\ dC(t)/dt = \alpha\cdot(C^*(t) - C(t)) \end{cases} \quad (1)$$

Where ηi is a per-module heterogeneity penalty factor derived from the measured flow deviation of module i relative to the rack average (Figure 1), δi(t) ∈ {0,1} is the binary valve state, and Rmref is the clean-membrane reference resistance. The system is integrated via the fourth-order Runge–Kutta (RK4) method with a fixed time step matched to the sensor sampling rate.

Parameter calibration. The five free parameters (βcake, βirr, ηi, α, and the chemical attenuation constants 0.04/0.05) were calibrated on a dedicated training partition of the 150-hour window (the first 80% of the sequence, approximately 120 hours) by minimizing the mean squared error between simulated and measured TMP trajectories using a grid-search procedure. Model fidelity was then assessed on the held-out chronological test partition (remaining 20%, approximately 30 hours). This chronological split was chosen deliberately to respect the temporal ordering of the data and avoid data leakage. It is acknowledged that, given the single-site, single-window nature of this study, the reported metrics are best interpreted as indicative performance bounds rather than statistically generalized results; cross-validation over multiple independent operational windows and sensitivity analysis over the parameter space constitute important directions for future work.

3. Reference Rack-Average Strategy

To enable a rigorous comparison, the baseline "rack-average" control strategy is defined operationally as follows: a global backwash event is triggered simultaneously across all modules whenever the rack-average hydraulic resistance R(t) = (1/n)Σi=1n Ri(t) exceeds a fixed threshold Rth; the coagulant dose Cref is held constant at the dataset mean dosing level. This corresponds to the standard industrial protocol described in [12], serving as the benchmark against which the proposed asynchronous, module-level strategy is evaluated.

4. Hybrid TCN and Asynchronous Control

The 150-hour dataset was augmented with simulator outputs—module-specific resistances (Ri) and optimal coagulant doses—yielding a hybrid dataset that translates lagging global indicators into precise module-level targets. A multi-output TCN [18] is trained on this enriched dataset to simultaneously forecast Ri and coagulant demand over a 5-minute horizon, using dilated causal convolutions that prevent future information leakage and capture long-range dependencies through residual connections.

The TCN architecture was deliberately favored over recurrent alternatives such as Long Short-Term Memory (LSTM) networks [19]. Unlike LSTMs, which process sequences iteratively and incur substantial computational overhead over long horizons, TCNs exploit dilated causal convolutions that can be trained in parallel, while stacked convolutional layers with residual connections maintain a wide, stable receptive field without the vanishing-gradient issues that can affect deep recurrent models—a property well suited to relating transient, high-frequency hydraulic signals to the slower cumulative dynamics of irreversible fouling.

Concretely, the multi-output TCN comprises four stacked residual blocks with dilation rates $d \in \{1, 2, 4, 8\}$, each containing two causal convolutional layers (32 filters, kernel size 3, L2 weight regularisation of 10^{-4}) followed by batch normalisation, Swish activation, and spatial dropout (rate 0.15); a 1×1 convolution matches channel dimensions on the skip connection where required. The stacked blocks are followed by global average pooling, a dense layer (32 units, Swish activation) with dropout (rate 0.2), and a final linear output layer producing the four target variables (Ri for each module and the coagulant dose) simultaneously. The network takes as input an 11-feature sequence—the three module flow rates, feed temperature, the three module resistances and their first-order time derivatives, and the coagulant dose—over a 90-step input window. It forecasts the same four targets 15 steps (5 minutes) ahead. The coagulant-dose signal was smoothed with a 20-step rolling mean before training to suppress pump-induced measurement noise, and all inputs and targets were independently min–max scaled. The model was trained for up to 30 epochs (batch size 32, 20% validation split) using the Adam optimizer (learning rate 5×10^{-4}) with a Huber loss, with early stopping (patience 8, best weights restored) and a learning-rate reduction on plateau (factor 0.5, patience 4). Hyperparameters (filter count, kernel size, dilation schedule, dropout rates, learning rate) were set empirically based on preliminary experimentation rather than through an automated search; systematic hyperparameter optimization is left for future work.

TCN forecasts drive an asynchronous module-level intervention logic: a targeted backwash alert is issued exclusively for the unit predicted to breach its individual critical resistance threshold within the 5-minute window. At the same time, the global coagulant dose is updated in real time based on evolving heterogeneity in resistance. The model was trained on the first 70% of the enriched sequence and evaluated on the chronologically held-out 30%.

III. RESULTS AND DISCUSSION

1. Backwash Mitigation and Chemical Optimization

By monitoring individual Ri trajectories, the asynchronous framework reduced the total number of backwash events from 328 (reference rack-average strategy, as defined in Section II-3) to 208, eliminating 120 redundant interventions. The evolution of coagulant dosage highlights the critical flaw of global-average reactive controllers: severe dosing volatility driven by lagging aggregate metrics. As shown in Figure 2, the proposed simulator yields a highly stable dosing profile and a projected 75.99% reduction in coagulant consumption over the evaluated 150-hour window. These figures are derived from a simulation-enriched evaluation on a single public dataset; they should be interpreted as scenario-specific estimates rather than guaranteed operational outcomes.

A mathematical wear-estimation submodel was used to project membrane lifespan. Polymer UF membranes degrade through two cumulative mechanisms: (i) mechanical fatigue from cyclic hydraulic stress, quantified as $W_{mec} = \Sigma_k \Delta P_k^2 \cdot \tau_k$, where ΔP_k and τ_k are the pressure amplitude and duration of backwash event k; and (ii) chemical degradation from coagulant overexposure, quantified as $W_{chem} = \int_0^T C(t)\, dt$. A composite wear index $W = w_1 W_{mec} + w_2 W_{chem}$ (with empirical weights w_1 and w_2 taken from the membrane aging literature) is

compared across strategies. By averting 120 backwash cycles and curtailing coagulant exposure by ~76%, the proposed strategy reduces W sufficiently to project an estimated lifespan extension of 60.70% under the studied conditions.

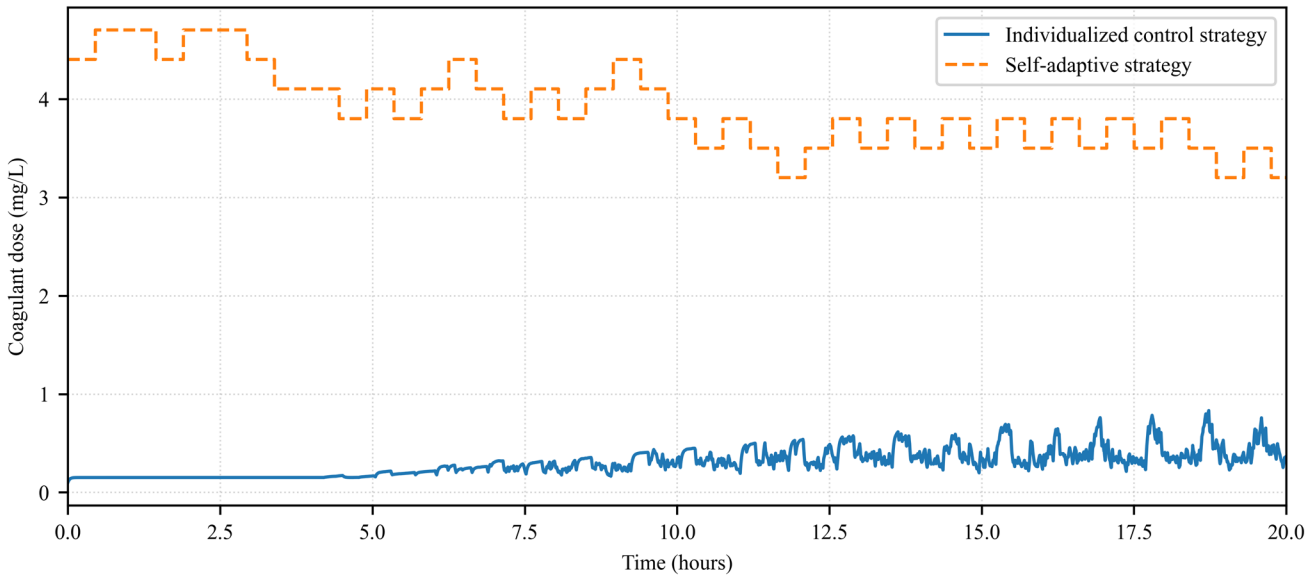


Figure 2. Zoomed comparison of coagulant dosing profiles (first 20 hours shown for clarity): the proposed approach achieves stable, demand-driven dosing versus the volatile reference strategy.

2. Multi-Objective Optimization and Pareto Trade-off

By averting 120 unnecessary backwash cycles, energy consumption dropped from 5.40×106 kWh to 3.66×106 kWh, a 32.16% saving under the simulated conditions. The filtrate purity index, modeled as:

$$\text{Quality Index (\%)} = \max(0,\ 100\times(1 - R_{irr} / R_{irr,crit})) \quad (2)$$

uses Rirr, crit = 0.6, a threshold calibrated on the dataset training partition to correspond to the onset of measurable permeate quality degradation, consistent with the 95% purity requirement reported in the reference industrial protocol [12]. The proposed strategy maintained a stable quality index superior to the rack-average approach throughout the evaluation period, despite a minor ~3.7% reduction in total volumetric production. This defines a clear Pareto trade-off: membrane health is prioritized over brute-force throughput, yielding a 36.59% reduction in immediate maintenance costs while preserving permeate quality above the operational threshold.

3. Predictive Performance of the TCN

The multi-output TCN was evaluated on the chronological held-out test partition for a t+5 min prediction horizon as shown in Figure 3. As reported in Table 1, the model achieves R2 = 0.9745 and RMSE = 0.0372 for coagulant dosing. For module-level resistances, R2 values range from 0.8267 to 0.8908, with all MAE values below 0.08. It is important to note that model evaluation was performed on a held-out chronological segment of a single, simulation-enriched operational trajectory. The reported metrics, therefore, reflect interpolation accuracy within

a continuous simulated regime rather than out-of-sample prediction on fully independent plant measurements; generalization to other sites or operating conditions requires additional validation.

It should also be noted that a systematic comparison against alternative sequence-learning baselines (e.g., LSTM-based predictors) and a formal statistical significance analysis (e.g., confidence intervals via bootstrap resampling across multiple training runs) were not performed in this proof-of-concept, single-window study; both are identified as necessary steps toward a fully validated benchmark, alongside the cross-validation and sensitivity analyses already noted in Section II-2.

Notably, Module 2 exhibits a consistently higher error margin ($R2 = 0.8267$, $MAE = 0.0748$) compared to Modules 1 and 3. This is physically consistent with the greater hydrodynamic deviation of Module 2 from the rack mean: its flow distribution departs more significantly from the uniform assumption embedded in the simulator's nominal parameters, introducing additional stochasticity in its fouling trajectory that the TCN must resolve with less calibration support. In a rack-average model, this module-specific anomaly would be invisible, absorbed into a global mean; the proposed framework, by processing each module individually, ensures a localized response rather than a premature rack-wide backwash.

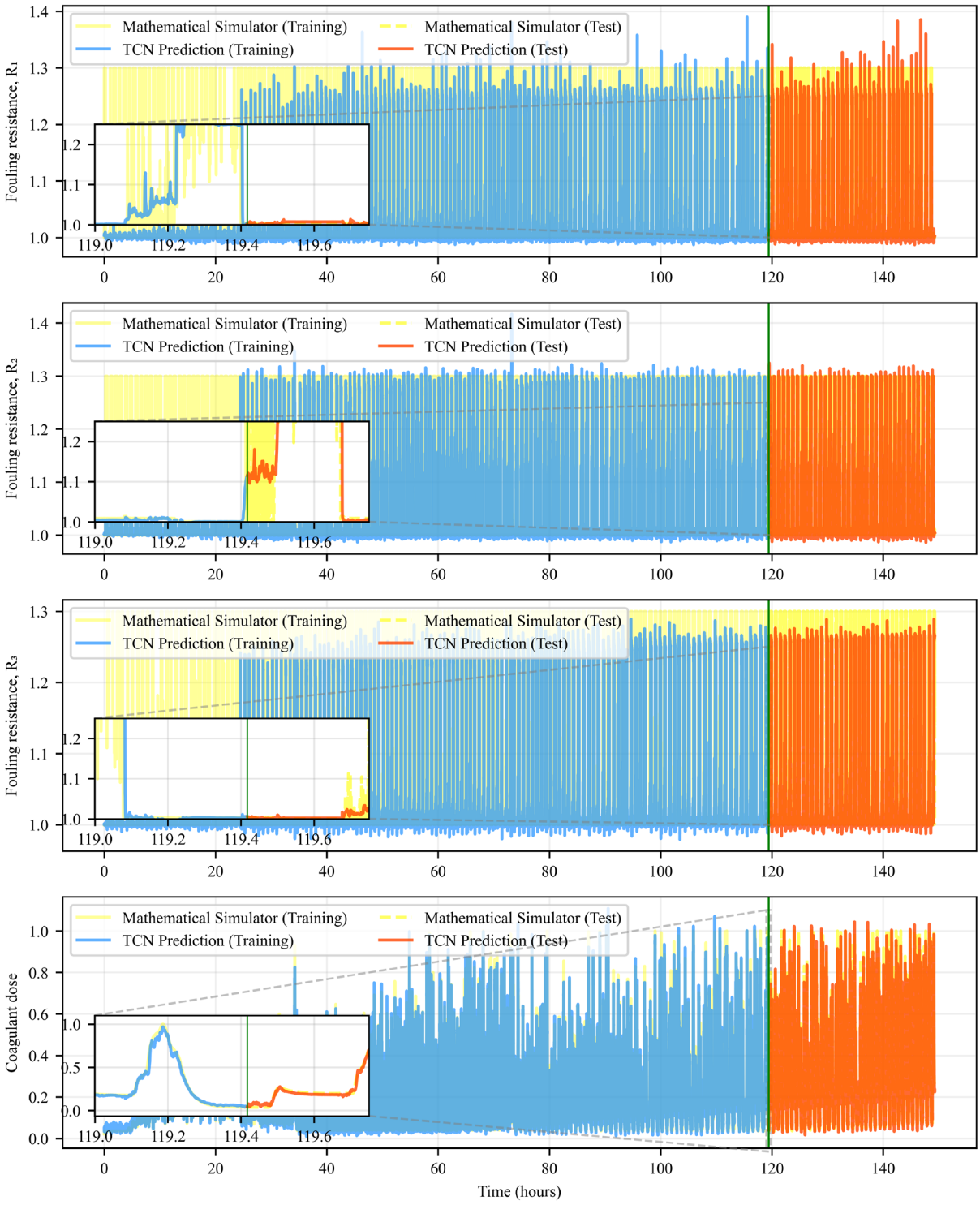


Figure 3. Predictive performance of the multi-output TCN on the chronological held-out test partition: individual module resistance (Ri) and optimal coagulant dose forecasted at a t+5 min horizon.

Table 1. Statistical performance metrics of the multi-output TCN on the held-out chronological test partition (5-minute prediction horizon).

Predicted Variable	**MAE**	**RMSE**	**R^2**
R (Module 1)	0.0559	0.1296	0.8908
R (Module 2)	0.0748	0.1777	0.8267
R (Module 3)	0.0583	0.1289	0.8891
Coagulant Dose	0.0294	0.0372	0.9745

IV. CONCLUSION

This work presents a hybrid physics-informed predictive supervision framework for UF pre-treatment in seawater desalination. By addressing the limitation of rack-average control—the masking of hydrodynamic heterogeneity across parallel modules—the proposed architecture shifts from synchronized reactive maintenance to asynchronous, module-level proactive intervention. A high-fidelity RK4-based mathematical simulator, with parameters calibrated on a 70% chronological training partition, combined with a multi-output TCN, enables 5-minute-ahead forecasting of individual module resistance and optimal chemical demand. Together, these three pillars—physics-based state estimation, module-level resolution, and TCN-driven forecasting—directly address the aggregation bias identified in the literature (Section I), converting hydrodynamic heterogeneity from an overlooked confound into an actionable control signal. Comparative simulation-based analysis against a formally defined rack-average baseline suggests the following potential improvements under the studied conditions:

- (i) an estimated 36.59% reduction in maintenance costs and 60.70% membrane lifespan extension, by eliminating 120 redundant backwash cycles and reducing coagulant overexposure;
- (ii) a 75.99% reduction in coagulant consumption and a 32.16% decrease in energy consumption (saving over 1.7 million kWh);
- (iii) strong predictive fidelity on the held-out test partition ($R^2 = 0.9745$ for dosing; $R^2 \approx = 0.89$ for module resistance), with a minor ~3.7% volumetric production trade-off.

These results are derived from a single-site, simulation-enriched study on a 150-hour window and should be interpreted as indicative upper bounds. Confidence intervals, sensitivity analyses over the model parameter space, and cross-validation over multiple independent operational windows are identified as essential next steps. As a non-invasive, software-driven solution compatible with existing SCADA and PLC environments, this framework constitutes a scalable foundation for plant-wide digital twin deployment and full-scale experimental validation on additional desalination sites.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Yoram Cohen et al. for making their UF pre-treatment dataset publicly available via the Dryad repository. The authors also thank ONEE (Water Branch, Morocco) for background information on industrial-scale UF and RO operations. The first author gratefully acknowledges the Center National pour la Recherche Scientifique et Technique (CNRST, Morocco) for financial support through the PASS doctoral scholarship.

REFERENCES

[1]. W. Musie and G. Gonfa, “Fresh water resource, scarcity, water salinity challenges and possible remedies: A review,” Heliyon, vol. 9, no. 8, p. e18685, 2023, doi: 10.1016/j.heliyon.2023.e18685.

[2]. W. Yu, N. Graham, and T. Liu, "Prevention of UF membrane fouling in drinking water treatment by addition of H2O2 during membrane backwashing," Water Research, vol. 149, pp. 394–405, 2019, doi: 10.1016/j.watres.2018.11.054.
[3]. J. Zhang et al., "Reduction of ultrafiltration membrane fouling by the pretreatment removal of emerging pollutants: A review," Membranes, vol. 13, no. 1, p. 77, 2023, doi: 10.3390/membranes13010077.
[4]. R.-S. Juang, H.-L. Chen, and Y.-S. Chen, "Membrane fouling and resistance analysis in dead-end ultrafiltration of Bacillus subtilis fermentation broths," Separation and Purification Technology, vol. 63, no. 3, pp. 531–538, 2008.
[5]. C. Wang et al., "Membrane fouling mechanism in ultrafiltration of succinic acid fermentation broth," Bioresource Technology, vol. 116, pp. 366–371, 2012.
[6]. N. I. Ishak, D. C. J. Chieh, L. W. Jiun, and O. B. Seng, "Dynamic flux modeling of a modular ultrafiltration unit with depleting hydrostatic pressure and its fouling study," Materials Today: Proceedings, 2023.
[7]. J. J. Wu, "Improving membrane filtration performance through time series analysis," Discover Chemical Engineering, vol. 1, 2021.
[8]. W. H. Abuwatfa, N. AlSawaftah, N. Darwish, W. G. Pitt, and G. A. Husseini, "A review on membrane fouling prediction using artificial neural networks (ANNs)," Membranes, vol. 13, no. 7, p. 685, 2023.
[9]. D. J. Kovacs et al., "Membrane fouling prediction and uncertainty analysis using machine learning: A wastewater treatment plant case study," Journal of Membrane Science, vol. 660, p. 120817, 2022.
[10]. A. Golabi, A. Erradi, and H. Qiblawey, "Optimal operation of reverse osmosis desalination process with deep reinforcement learning methods," Applied Intelligence, vol. 54, no. 8, pp. 6333–6353, 2024.
[11]. S. Kherraf et al., "Forecasting of permeate conductivity using MLR and ANN methods of Boujdour seawater reverse osmosis desalination plant," Current Analytical Chemistry, vol. 19, no. 4, pp. 348–355, 2023.
[12]. L. X. Gao, H. Gu, A. Rahardianto, P. D. Christofides, and Y. Cohen, "Self-adaptive cycle-to-cycle control of in-line coagulant dosing in ultrafiltration for pre-treatment of reverse osmosis feed water," Desalination, vol. 401, pp. 22–31, 2017, doi: 10.1016/j.desal.2016.09.024.
[13]. Y. Zhou, B. Khan, H. Gu, P. D. Christofides, and Y. Cohen, "Modeling UF fouling and backwash in seawater RO feedwater treatment using neural networks with evolutionary algorithm and Bayesian binary classification," Desalination, vol. 513, p. 115129, 2021, doi: 10.1016/j.desal.2021.115129.
[14]. A. Abdulsalam et al., "A review of fouling prediction techniques in membrane desalination: From empirical to intelligent models," Desalination, vol. 616, p. 119363, 2025, doi: 10.1016/j.desal.2025.119363.
[15]. S. Saeedi Garakani and J. W. Chew, "Development of physics-informed machine-learning models to enhance understanding and prediction of membrane fouling," Journal of Membrane Science, vol. 728, p. 124133, 2025, doi: 10.1016/j.memsci.2025.124133.
[16]. F. J. Abdu et al., "Design of real-time hybrid nanofiltration/reverse osmosis seawater desalination plant performance based on deep learning application," Desalination, vol. 611, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0011916425003935
[17]. Y. Cohen, Y. Zhou, B. Khan, and H. Gu, "UF pre-treatment of seawater RO feedwater – performance data." Dryad, 2021. doi: 10.5068/D1310B.
[18]. S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." [Online]. Available: https://doi.org/10.48550/arXiv.1803.01271
[19]. C. B. Vennerød, A. Kjærran, and E. S. Bugge, "Long short-term memory RNN." [Online]. Available: https://doi.org/10.48550/arXiv.2105.06756

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Leakage-Aware and Explanation-Stable XAI for Short-Term Smart Home Appliance Energy Forecasting

Ahmad Wedyan
School of Digital Technologies
Manukau Institute of Technology
Auckland, New Zealand
ahmad.wedyan@manukau.ac.nz
ORCID: 0000-0002-7137-3520

David McCurdy
School of Digital Technologies
Manukau Institute of Technology
Auckland, New Zealand
david.mccurdy@manukau.ac.nz
ORCID: 0009-0003-7248-2299

Abstract

Smart home energy systems generate high-frequency Internet of Things (IoT) sensor data that can support short-term, one-step-ahead appliance energy forecasting. However, two methodological risks persist: temporal data leakage, which can inflate performance metrics, and the causal misinterpretation of explainable AI (XAI) outputs. This paper proposes a leakage-aware and explanation-stable machine learning framework and evaluates it on the UCI Appliances Energy Prediction dataset, which contains 19,735 observations at 10-minute intervals. Six model architectures and 13 feature group configurations were evaluated using five-fold TimeSeriesSplit cross-validation on the first 80% of the training data; the final 20% (3,943 rows) was reserved as an untouched prospective holdout. The selected Random Forest model, using 22 temporal and autoregressive features, achieved RMSE = 58.47 Wh and $R^2 = 0.5826$, outperforming the previous-value persistence baseline (RMSE = 66.23 Wh). A 1,000-sample approximate paired-bootstrap audit produced a 95% confidence interval for RMSE reduction of [4.98, 10.65] Wh. SHAP analysis showed that the selected model's predictive structure was dominated by recent appliance-consumption history. In this single-household, one-step-ahead setting, environmental IoT variables added limited incremental value once lag and rolling consumption features were included. Across five seed-specific model-and-sample runs, the top-10 SHAP feature sets achieved an average pairwise Jaccard similarity of 0.9273. A deployment-oriented ablation indicated that smart-meter-only lag and rolling features achieved training-period cross-validation performance close to the selected 22-feature configuration. All findings are framed as model-based associations rather than causal claims, supporting cautious use in energy-awareness dashboards.

Keywords: explainable AI, SHAP, smart home, IoT, short-term energy forecasting, leakage-aware validation, random forest, feature ablation, explanation stability, bootstrap confidence intervals, HEMs.

I. INTRODUCTION

The growth of IoT devices, smart meters, and home-automation systems has increased the availability of high-frequency environmental and energy-consumption data in residential settings. These data streams enable short-term, one-step-ahead appliance energy forecasting for home energy management systems, demand-side response, and consumer energy awareness dashboards [1], [2]. However, two methodological issues continue to limit reliability. Temporal leakage occurs when randomized validation allows future observations to influence model training or selection, producing overly optimistic performance estimates [3], [4]. Causal misinterpretation of XAI outputs is also a concern: SHAP values explain the behavior of the trained model but do not, by themselves, establish physical causal mechanisms [5], [6]. This paper proposes a leakage-aware and explanation-stable framework for short-term appliance energy forecasting, evaluated on the UCI Appliances dataset across 13 feature configurations and six model families. The framework combines leakage-safe autoregressive feature

construction, strict chronological validation, training-only TimeSeriesSplit model selection, operational baseline comparison, bootstrap uncertainty estimation, feature-domain ablation, and SHAP stability auditing. These components reduce the risk of temporally inflated performance estimates and overconfident interpretations of unstable or causally overstated XAI outputs. The study adopts a contextual coherence perspective in which model outputs are treated as context-dependent associations rather than causal mechanisms. Under this perspective, trustworthy forecasting requires consistency across temporal, procedural, operational, statistical, and explanatory contexts. This motivates the emphasis on leakage-aware validation, operational baselines, uncertainty estimation, and explanation-stability auditing, which together assess whether the model behaves coherently under the conditions in which it is trained, selected, evaluated, and interpreted. Four research questions guide the study. RQ1 examines how training only a chronological model selection affects final model choice and holdout interpretation. RQ2 evaluates the contribution of lag and rolling features relative to previous value persistence. RQ3 assesses whether historical consumption, temporal encodings, or environmental IoT variables dominate the model's SHAP interpreted structure. RQ4 examines whether global SHAP feature sets remain stable across repeated seed-specific model and sample runs. The contribution is an audited evaluation protocol for one-step-ahead appliance energy forecasting in settings where recent meter readings are available. The protocol integrates leakage-safe feature engineering, chronological model selection, operational baselines, feature-domain ablation, uncertainty estimation, spike-regime diagnostics, minimal-sensor analysis, and SHAP stability auditing. Tree-based models are evaluated because they perform strongly on medium-sized tabular datasets [7]; the evaluated model families include Random Forest [8], XGBoost [9], and LightGBM [10].

II. RELATED WORK

Short-term residential load forecasting has been widely studied using statistical and machine-learning methods [1], [2]. Tree-based methods remain strong benchmarks for medium-sized tabular datasets [7], while Random Forest [8], XGBoost [9], and LightGBM [10] provide established ensemble architectures. The UCI Appliances dataset established a widely used benchmark based on indoor environmental measurements and external weather data. In the original study, the strongest Gradient Boosting Machine model explained approximately 57% of the variance in the testing set [11]. SHAP and related XAI methods are increasingly used to interpret energy-forecasting models [5], [6], [12–15]. The original SHAP framework provides additive feature attributions for individual predictions [12], while TreeSHAP supports efficient explanation of tree-based models and aggregation from local to global model behavior [13]. Energy-forecasting studies have subsequently used Shapley-based explanations to identify influential model inputs [14], [15]. More generally, analytical conclusions can vary across modeling choices, motivating explicit robustness and stability checks [16]. SHAP values should therefore be interpreted as model-based attributions rather than physical causal effects [5], [6]. This study integrates leakage-aware validation, training-only selection, operational baselines, feature-domain ablation, uncertainty estimation, spike diagnostics, minimal-sensor analysis, and SHAP stability auditing into a single framework.

III. METHODOLOGICAL FRAMEWORK

The framework consists of five stages: data preparation, leakage-safe feature engineering, temporal validation, multi-model feature-group ablation, and SHAP-based explanation stability auditing. The aim is not to introduce a new forecasting algorithm but to evaluate short-term appliance energy forecasting under conditions that more closely reflect prospective smart home deployments. The evaluation structure follows contextual coherence. Temporal coherence is maintained through leakage-safe feature engineering using a one-step shift so that each feature vector at time t uses only information available at t−1. Procedural coherence is ensured through training

only chronological model selection using a 5-fold TimeSeriesSplit [17], which expands the training window and evaluates performance on the immediately following block. Operational coherence is assessed against previous-value persistence, a simple operational benchmark for one-step-ahead forecasting. Statistical coherence was assessed using 1,000 paired bootstrap resamples of the holdout set, with 95% intervals obtained from the 2.5th and 97.5th percentiles. Because holdout observations are temporally dependent, these intervals are treated as an approximate uncertainty audit rather than formal time-series confidence bounds. Explanatory coherence was evaluated using TreeSHAP [12], [13] across five model-and-sample runs with seeds 1, 7, 21, 42, and 99. For each run, TreeSHAP was applied to up to 2,000 randomly sampled holdout observations, and stability was assessed using pairwise Jaccard similarity between the top-10 SHAP feature sets. Because each seed affected both model fitting and SHAP sampling, the audit measures combined model and sample stability. The study uses the UCI Appliances Energy Prediction dataset [11], containing 19,735 observations recorded at 10-minute intervals over approximately 4.5 months in a low-energy house in Belgium. After constructing leakage-safe features, 19,711 usable observations remained. The first 15,768 observations were used for training and TimeSeriesSplit cross-validation, and the final 3,943 observations were reserved as an untouched chronological holdout. The variables rv1 and rv2 were removed because they are random variables without operational meaning. Three additional clarifications prevent common interpretation errors: Visibility is an external weather-station variable, lights measure lighting energy use rather than occupancy, and the dataset contains no direct occupancy sensor. The dataset, therefore, supports controlled evaluation of appliance energy forecasting and feature-domain contribution but cannot directly represent occupancy or individual appliance events. Historical target features were constructed using shift(1) to prevent leakage. The selected 22-feature set comprised 11 temporal features (hour, day of week, weekend, month, minutes since midnight, and their relevant sine/cosine encodings), six appliance lags (1, 2, 3, 6, 12, and 24 steps), and five shifted rolling statistics (means over 3, 6, and 12 steps; standard deviations over 6 and 12 steps). Because the feature set includes a 24-step lag and shifted rolling windows, the first 24 observations were dropped. This design assumes an online, one-step-ahead forecasting scenario in which previously observed smart-meter readings are available before prediction of the next interval. This is operationally realistic for a HEMS but limits the framework to next-interval forecasting using recent consumption history. Strict chronological partitioning was used, with the first 80% of observations allocated to training and cross-validation and the final 20% reserved as a prospective holdout. No future holdout values were used to construct features at an earlier timestamp. Model selection relied exclusively on the mean TimeSeriesSplit RMSE, and the holdout was evaluated only after the selection rule was applied. This design directly addresses RQ1 by preventing retrospective holdout-based model selection. XGBoost achieved the lowest holdout RMSE of 57.74 Wh, but Random Forest was retained as the selected model because the prespecified selection rule relied solely on training-period cross-validation. The leading cross-validation configurations were tightly clustered: Random Forest achieved RMSE = 64.55 Wh, HistGradientBoosting 64.68 Wh, LightGBM 64.74 Wh, and XGBoost 64.86 Wh. These results indicate that the leading architectures performed similarly under the fixed evaluation settings. Six model families were evaluated using fixed hyperparameters: Random Forest, XGBoost, LightGBM, HistGradientBoosting, ExtraTrees, and Ridge regression. The selected Random Forest used 80 trees, max_depth = 14, min_samples_leaf = 2, and max_features = "sqrt". Thirteen feature-group configurations produced 78 model–feature combinations. The baselines were the training mean, a seasonal day/time mean, and previous-value persistence. RMSE was the primary metric, with MAE and R^2 reported as secondary metrics. For the primary explanation, TreeSHAP was applied to 2,000 holdout observations using seed 42. High Jaccard similarity across the repeated runs indicates robustness of the global top-10 feature set, but it does not establish causal validity or complete explanation stability.

IV. EXPERIMENTAL RESULTS

Table 1 reports the holdout performance of the operational baselines, the selected model, and the XGBoost sensitivity result. The selected Random Forest achieved RMSE = 58.47 Wh, MAE = 26.31 Wh, and R^2 = 0.5826, outperforming all three operational baselines. The most important comparator is previous-value persistence, which achieved an RMSE of 66.23 Wh and an R^2 of 0.4645. The selected model, therefore, provides forecasting value beyond a simple lag-1 strategy. Although XGBoost achieved a slightly lower holdout RMSE of 57.74 Wh, it was not selected because model selection was based exclusively on training-period TimeSeriesSplit cross-validation. It is consequently reported as a sensitivity result rather than as the final model. Together with the tightly clustered cross-validation results, its small holdout advantage suggests that leakage-safe lag and rolling features were more influential than minor differences among the leading model families.

Table 1. Holdout performance of baselines and selected models

Model/baseline	**RMSE (Wh)**	**MAE (Wh)**	**R^2**	**Note**
Dummy mean	90.53	52.45	-0.0005	Training mean
Seasonal day/time mean	86.57	47.68	0.0850	Hour/day mean
Previous-value persistence	66.23	26.53	0.4645	Lag-1 baseline
Random Forest, selected	58.47	26.31	0.5826	CV-selected
XGBoost, sensitivity	57.74	25.46	0.5930	Holdout sensitivity

Table 2 reports approximate paired-bootstrap 95% intervals based on 1,000 resamples. The selected Random Forest reduced holdout RMSE from 66.23 Wh to 58.47 Wh, an absolute reduction of 7.76 Wh or 11.72% relative to persistence. The RMSE reduction interval of [4.98, 10.65] Wh remained positive, suggesting that the improvement was not dependent on a fragile point estimate under this approximate row-resampling audit.

Table 2. Bootstrap uncertainty for the selected model and improvement over the persistence baseline

Quantity	**Point Estimate**	**95% CI**
Selected model RMSE Wh	58.47	[53.90, 63.15]
Selected model R^2	0.5826	[0.537, 0.626]
RMSE reduction vs lag-1 Wh	7.76	[4.98, 10.65]
RMSE reduction vs lag-1 %	11.72%	[7.44%, 16.00%]

Published results obtained using the same UCI dataset provide useful context. Candanedo et al. reported a test set RMSE of 66.65 Wh, an MAE of 35.22 Wh, and an R^2 of 0.57 for their Gradient Boosting Machine model [11]. Assadian and Assadian reported RMSE = 59.61 Wh, MAE = 26.62 Wh, and R^2 = 0.66 for a tuned Extra Trees model [18]. The present Random Forest achieved RMSE = 58.47 Wh, MAE = 26.31 Wh, and R^2 = 0.5826. These values indicate competitive, but not directly superior, predictive performance. The studies used different predictors, feature-engineering procedures, tuning strategies, forecasting tasks, and data-partitioning protocols. In particular, the present study evaluates one-step-ahead forecasting using leakage-safe autoregressive features, training-only chronological selection, and an untouched final chronological holdout. The published values should therefore be treated as contextual benchmarks rather than a controlled head-to-head ranking. Feature groups without an autoregressive consumption history produced cross-validation RMSE values above approximately 90 Wh. In contrast, configurations containing lag and rolling history achieved RMSE values of approximately 64.5–66.9 Wh across the evaluated architectures. The selected Random Forest using temporal,

lag, and rolling features achieved the lowest mean cross-validation RMSE of 64.55 ± 4.87 Wh, with MAE = 31.20 ± 4.20 Wh and $R^2 = 0.598 \pm 0.037$. These findings address RQ2. In this one-step-ahead benchmark, recent appliance consumption history was the dominant predictive signal, whereas environmental IoT variables without autoregressive consumption history were insufficient for competitive forecasting. Figure 1 presents the best cross-validated RMSE for each feature group across the 78 model–feature combinations. The complete fold-level cross-validation results are omitted for brevity.

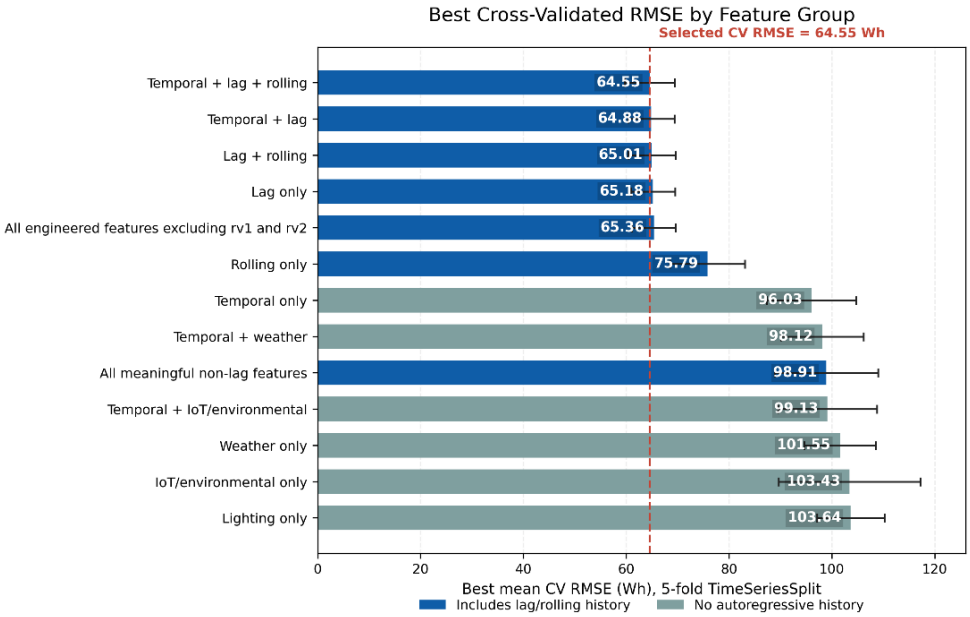


Figure 1. Best training-period CV RMSE by feature group using five-fold TimeSeriesSplit; error bars show fold-level standard deviation.

Table 3 reports holdout error by actual consumption band. The model performed well during low- and moderate-consumption periods, but errors increased sharply above 300 Wh, indicating systematic underprediction during sudden high-load events.

Table 3. Holdout error by actual consumption band

Actual Band	Count	MAE (Wh)	Underprediction	Interpretation
0–100 Wh	3,120	14.50	33.37%	Stable low-use periods
100–300 Wh	652	41.56	40.80%	Moderate routine-use error
>300 Wh	171	183.72	85.38%	Severe spike underestimation

The >300 Wh band was used as a diagnostic high-consumption regime rather than as a model-selection criterion. It represents an upper-tail subset of the holdout observations. For this band, MAE reached 183.72 Wh, and 85.38% of observations were underpredicted, indicating a structural limitation in the available feature set and in the one-step-ahead task design. Sudden high-load events are difficult to anticipate from lag and rolling consumption features alone, particularly in the absence of direct occupancy or appliance-event sensing.

V. EXPLAINABILITY AND STABILITY ANALYSIS

TreeSHAP was applied to the selected Random Forest after confirming that it outperformed both the dummy-mean and previous-value persistence baselines on the holdout set. For the primary explanation, SHAP values were calculated using a random sample of 2,000 holdout observations selected with seed 42. Figure 2 presents the global SHAP bar summary. The five highest-ranked features were all autoregressive consumption-history variables: Appliances_lag_1 with mean absolute SHAP = 21.58 Wh, Appliances_rolling_mean_3 with 10.54 Wh, Appliances_rolling_std_6 with 6.98 Wh, Appliances_rolling_std_12 with 5.14 Wh, and Appliances_rolling_mean_6 with 4.85 Wh. The remaining top-10 features were Appliances_lag_2, hour_cos, Appliances_rolling_mean_12, minutes_since_midnight_sin, and Appliances_lag_3. No environmental IoT variable appeared in the top 10, directly addressing RQ3.

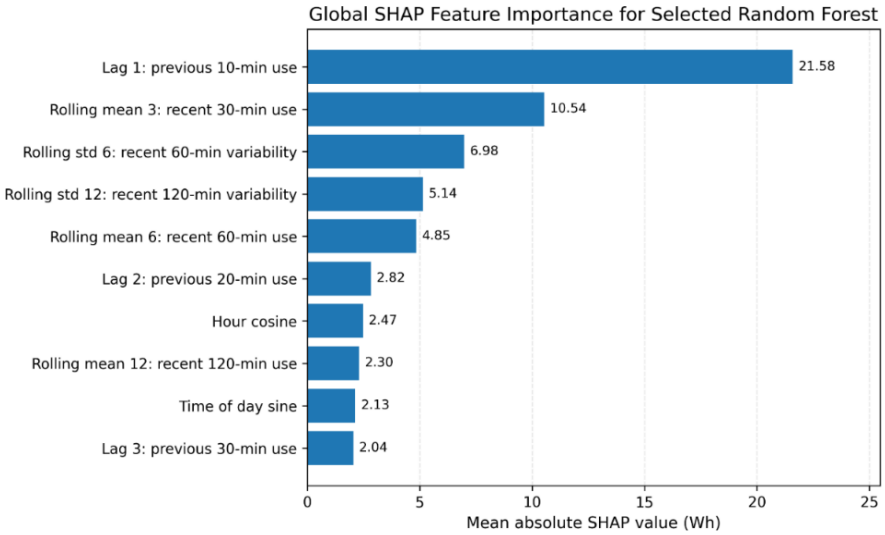


Figure 2. Global SHAP feature importance for the selected Random Forest, showing the dominance of recent appliance-consumption history.

The SHAP ranking is consistent with the feature-group ablation: environmental-only configurations performed substantially worse than configurations containing lag and rolling history, and no environmental IoT variable appeared among the top-10 SHAP features. All SHAP findings are interpreted as model-based associations rather than causal effects. The following rule was applied when translating SHAP outputs into possible dashboard feedback: "Feature X is associated with higher or lower predicted appliance energy use in the trained model. This may reflect household routines or conditions that co-occur with appliance use, but the model does not demonstrate that changing X alone will alter actual energy consumption."

For example, the association between Appliances_lag_1 and predicted energy reflects short-term consumption persistence: appliances operating during the previous 10-minute interval may remain active during the following interval. This is evidence of temporal autocorrelation in the model's predictive behavior, not evidence of a causal intervention pathway. Figure 3 presents pairwise Jaccard similarities among the top-10 SHAP feature sets produced by the five seed-specific model-and-sample runs. Across the ten unique seed pairs, six pairs among seeds 1, 7, 21, and 42 produced identical top-10 feature sets. The four pairs involving seed 99 produced a Jaccard similarity of 0.818, corresponding to one feature difference at the top-10 boundary. The resulting average

pairwise Jaccard similarity was 0.9273. This result addresses RQ4 by demonstrating high stability at the feature-set level across the repeated runs. However, the audit does not isolate model-initialization effects from explanatory-sample variation, nor does it assess rank-order stability, attribution-magnitude stability, or local-explanation stability. The high feature-set overlap supports cautious use of the explanations for model-level energy-awareness feedback. It does not establish real-world causal validity or complete trustworthiness in deployment. Instead, it indicates that the model's broad global explanation pattern remained similar across repeated seed-specific model-and-sample runs, providing a more defensible basis for model-level dashboard summaries than a single-run explanation. User-facing interpretation would still require deployment and user-comprehension testing.

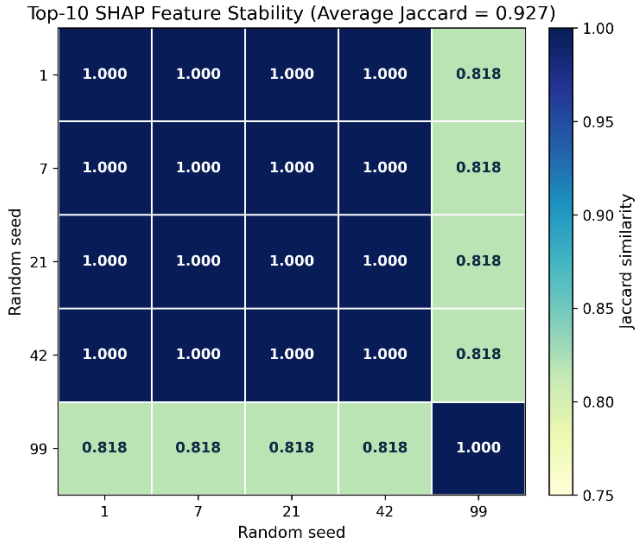


Figure 3. Top-10 SHAP feature-set stability across five seed-specific Random Forest models and sample runs. All pairs not involving seed 99 have identical top-10 feature sets; the four pairs involving seed 99 have Jaccard similarity of 0.818. The average pairwise similarity is 0.9273.

VI. DISCUSSION

Table 4 presents a deployment-oriented ablation based on training-period cross-validation. Each row reports the best model obtained for the corresponding feature group. Random Forest was the best model for the smart-meter-only and smart-meter-plus-time configurations, Ridge was the best model for the environmental-only configuration, and XGBoost was the best model for the full engineered feature set. The table provides preliminary deployment guidance and does not change the training-only selection rule or the reported Random Forest holdout result.

Table 4. Minimal sensor deployment ablation

Deployment setup	Feature group	Features	CV RMSE (Wh)	CV R^2
Smart meter only	Lag + rolling	11	65.01	0.592

Deployment setup	Feature group	Features	CV RMSE (Wh)	CV R^2
Smart meter + time	Temporal + lag + rolling	22	64.55	0.598
Environmental only	IoT/environmental only	19	103.43	-0.029
Full engineered setup	All engineered features	47	65.36	0.588

The smart-meter-only configuration using 11 lag and rolling features achieved a CV RMSE of 65.01 Wh, close to the 22-feature temporal, lag, and rolling configuration, which achieved 64.55 Wh. Environmental-only features performed poorly, with a CV RMSE of 103.43 Wh and a negative R^2. The full 47-feature engineered configuration achieved CV RMSE = 65.36 Wh and did not improve on the selected 22-feature configuration. These results suggest that smart-meter lag and rolling features may provide an efficient basis for low-cost HEMS forecasting. However, the compact feature set must still be evaluated on an untouched holdout or external households. The compact configuration may support lightweight edge deployment. This remains contingent on validating computational requirements, latency, memory usage, handling of missing readings, and operational robustness. For energy-awareness dashboards, SHAP outputs should be framed as non-causal model explanations. Dashboards should also disclose the prediction limitation during high-load spikes above 300 Wh, where the selected model systematically underpredicts demand. Any deployed forecasting model should be evaluated against simple operational baselines such as previous-value persistence. Taken together, the ablation, spike-regime, and explanation-stability findings support the contextual-coherence interpretation. The model behaves coherently when recent consumption history is available, but becomes less reliable during abrupt high-load events not captured by occupancy or appliance-event variables. The stable global SHAP feature set supports reproducible model-level interpretation. At the same time, the minimal-sensor ablation indicates that most of the predictive value comes from smart-meter history rather than from additional environmental sensors. Limitations include the single-household dataset, absence of occupancy or appliance-event sensing, and reliance on an online one-step-ahead setting. SHAP explains model behavior rather than causal mechanisms, and the bootstrap does not fully preserve temporal dependence. Hyperparameters were constrained to ensure reproducibility rather than through exhaustive optimization. Real-world deployment requires appropriate consent, privacy controls, and secure handling of recent meter-reading windows, especially important in single-household datasets.

VII. CONCLUSION

This paper presented a leakage-aware and explanation-stable framework for short-term smart-home appliance energy forecasting. Using strict chronological validation and operational baselines, the selected Random Forest achieved RMSE = 58.47 Wh and R^2 = 0.5826 on an untouched holdout set, improving over previous-value persistence by 7.76 Wh in RMSE. Feature-group ablation and SHAP analyses showed that recent appliance consumption history dominated the prediction in this single-household, one-step-ahead setting. At the same time, environmental IoT variables added limited incremental value once lag and rolling features were included. Spike-regime analysis showed severe underprediction above 300 Wh, highlighting the need for occupancy-aware or appliance-event sensing. SHAP stability analysis supported cautious model-level explanation, provided outputs are treated as associations rather than causal claims. Future work should evaluate the framework across multi-household datasets, different climatic regions, and longer temporal horizons to assess generalisability and robustness.

REFERENCES

[1]. [1] K. Amasyali and N. M. El-Gohary, “A review of data-driven building energy consumption prediction studies,” *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1192–1205, Jan. 2018, doi: 10.1016/j.rser.2017.04.095.

[2]. [2] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," Renew. Sustain. Energy Rev., vol. 74, pp. 902–924, Jul. 2017, doi: 10.1016/j.rser.2017.02.085.

[3]. [3] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating time series forecasting models: an empirical study on performance estimation methods," Mach. Learn, vol. 109, no. 11, pp. 1997–2028, Nov. 2020, doi: 10.1007/s10994-020-05910-7.

[4]. [4] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice. OTexts, 2021. [Online]. Available: https://otexts.com/fpp3/

[5]. [5] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[6]. [6] Z. Chen, F. Xiao, F. Guo, and J. Yan, "Interpretable machine learning for building energy management: A state-of-the-art review," Adv. Appl. Energy, vol. 9, p. 100123, Feb. 2023, doi: 10.1016/j.adapen.2023.100123.

[7]. [7] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.08815

[8]. [8] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[9]. [9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[10]. [10] G. Ke et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017.

[11]. [11] L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data-driven prediction models of energy use of appliances in a low-energy house," Energy Build., vol. 140, pp. 81–97, Apr. 2017, doi: 10.1016/j.enbuild.2017.01.083.

[12]. [12] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[13]. [13] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[14]. [14] J. Moon, S. Rho, and S. W. Baik, "Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with Shapley values," Sustain. Energy Technol. Assessments, vol. 54, p. 102888, Dec. 2022, doi: 10.1016/j.seta.2022.102888.

[15]. [15] B. Devanathan, K. Jnana Varshitha, L. Pavan Kumar, S. A. Lakshmanan, and N. Krishna Prakash, "Explainable AI Framework Using XGBoost With SHAP and LIME for Multi-Scale Household Energy Forecasting," IEEE Access, vol. 13, pp. 149750–149764, 2025, doi: 10.1109/ACCESS.2025.3602673.

[16]. [16] S. J. Bell, O. P. Kampman, J. Dodge, and N. D. Lawrence, "Modeling the Machine Learning Multiverse," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2206.05985

[17]. [17] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1201.0490

[18]. [18] C. F. Assadian and F. Assadian, "Data-Driven Modeling of Appliance Energy Usage," Energies, vol. 16, no. 22, p. 7536, Nov. 2023, doi: 10.3390/en16227536.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Work-in-Progress Cross-Case Analysis of Three VR Exposure Prototypes

Marina Knyazhansky
Software Engineering Department
SCE – Shamoon College of Engineering
Israel
marinak@sce.ac.il
ORCID: 0000-0003-2928-7888

Abstract
Cost, safety concerns, unpredictability, and access may limit real-world exposure to specific phobias. Virtual reality (VR) provides a controllable and repeatable alternative. This work-in-progress paper presents second-generation VR systems for fear of cats, heights, and flying, and analyzes them as engineering artifacts. A preliminary cross-case comparison examines exposure hierarchy, intensity parameters, interaction and safety logic, and data requirements. Each system implements comprehensive VR event logging, smartwatch integration, temporal synchronization, and backend session histories. At the same time, exposure logic remains phobia-specific: object-based for cats, spatial for heights, and event-sequence for flying. From these cases, we derive a preliminary framework separating common session, logging, synchronization, and backend functions from stimulus-specific exposure logic. A recommendation mechanism is under development, and a formal usability study is planned. No clinical efficacy study was conducted; the contribution is an implemented engineering foundation for later expert review, usability and feasibility evaluation, and clinically supervised research.

Keywords: virtual reality; specific phobias; exposure systems; human-computer interaction; software engineering; interaction design

I. INTRODUCTION

Specific phobias involve persistent fear and avoidance of particular objects or situations. Exposure-based interventions involve structured encounters with feared stimuli, but in vivo exposure may be costly, difficult to organize, or difficult to control safely [1],[2]. VR offers repeatable, configurable exposure in which distance, movement, elevation, openness, sound, timing, and behavior can be controlled and logged. Evidence supports VR exposure, particularly over passive controls, but does not show consistent superiority over active or in-vivo approaches [3-5]; its engineering value therefore also lies in control, repeatability, interaction, and measurement. Designing such systems requires translating exposure principles into software requirements for intensity, progression, control, safety, and logging. These requirements depend on the feared experience: animal exposure centers on a dynamic agent, height exposure on spatial perception and navigation, and flight exposure on a timed multimodal sequence. We compare three student-developed VR prototypes for cats, heights, and flying as object-based, spatial, and event-sequence systems. This is an engineering abstraction, not a clinical taxonomy, and the prototypes are not presented as validated treatments. The study asks how graded exposure can be translated into VR requirements (RQ1), which functions are shared or phobia-specific (RQ2), and which reusable guidelines support progression, control, safety, and behavioral data collection (RQ3). The contribution is a cross-case comparison and preliminary framework linking exposure structure to configurable parameters, interaction and safety logic, and behavioral measures.

II. BACKGROUND AND RELATED WORK

1. Exposure Principles and Current Evidence

Exposure uses structured encounters with feared and avoided stimuli. Meta-analytic evidence shows substantial symptom and approach-behavior improvements [1], while inhibitory-retrieval models emphasize testing threat expectations rather than requiring anxiety to decline continuously [2]. A technological hierarchy should therefore manipulate meaningful threat-related variables rather than merely increase visual realism. VR exposure is a well-developed application area [3]. Outcomes are generally better than those of passive controls, but advantages over active conventional or in vivo interventions are inconsistent [4],[5], cautioning against equating technological sophistication with clinical superiority. Recent studies report benefits across several phobias [6], including heights [7] and flying [8]; evidence for cat phobia is more limited, though it includes a randomized study [9]. Together, they support feasibility while illustrating that different phobias require distinct mechanisms of stimulus control and interaction.

2. Technical Design, User Experience, and Research Gap

VR exposure systems vary in display, tracking, navigation, interaction, stimulus behavior, progression, and user or supervisor control. A review of 70 studies found inconsistent technical reporting, limiting reproducibility and comparison [10]; implementations range from passive 360-degree content to interactive 3D environments with different locomotion and progression schemes. User control and comfort are central HCI concerns. Users may need to pause or terminate exposure without controlling the feared event itself. Usability evaluation remains heterogeneous [11-14], and cybersickness may overlap with anxiety responses; navigation, visual motion, frame stability, and duration should therefore be designed to minimize unrelated discomfort [12]. Programmable environments can record duration, completion, interruptions, position, proximity, and interactions, while wearable biosensors add physiological data [13]. Because logging and technical reporting are not yet standardized [10],[13], there is limited guidance on which software functions can be reused across phobias and which must remain stimulus-specific. This work addresses the engineering-HCI gap.

III. RESEARCH AND DEVELOPMENT METHOD

1. Study Design and Project Context

The study used an artifact-centered qualitative comparison of three functional VR prototypes for cats, heights, and flying, developed as final-year software engineering projects. They were analyzed as engineering artifacts; treatment efficacy and clinical outcomes were outside the scope of this analysis. All prototypes were developed in Unity 6 for Meta Quest 3S and shared six requirements: graded progression, repeatable stages, control over participation, predictable progression, clear stopping or withdrawal, and behavioral event logging, while retaining stimulus-specific exposure logic.

2. Second-Generation Architecture and Multimodal Data Pipeline

Second-generation versions add complete VR event logging and smartwatch integration. Wearable data are timestamped and synchronized with VR events so both streams can be represented on a unified session timeline. Synchronized data is stored on a backend that provides access to the session history. Logging, wearable integration, synchronization, and backend history have been implemented; a recommendation mechanism is under development, and a formal usability study is planned. These functions provide preliminary technical evidence that one data-management pattern can support the three exposure categories without replacing their phobia-specific logic.

3. Translation and Comparative Analysis

For each system, exposure principles were mapped to a hierarchy, controllable parameters, VR interactions, progression, safety rules, and logging. Cat exposure used agent visibility, distance, movement, orientation, and behavior; height exposure used elevation, depth, openness, edge proximity, barriers, and movement; flight exposure used coordinated cabin, taxiing, takeoff, flight, and turbulence events. Each artifact was examined independently and then compared across seven dimensions: exposure structure and unit, controllable parameters, interaction and progression logic, user control, safety, behavioral logging, and primary engineering challenge. Shared features formed the common design core; differences were classified as object-based, spatial, or event-sequence, producing the framework in Section V. The comparison was qualitative and based on implemented artifacts and project documentation; no statistical analysis was appropriate because there was no controlled participant sample or common outcome set. The cat-system developer's five-session use was treated only as non-independent formative feedback, not as evidence of usability or efficacy. Future participant studies require predefined measures, supervision, consent, secure data handling, and ethics approval.

IV. VR EXPOSURE PROTOTYPES

The prototypes share the common requirements and second-generation data functions above, but differ in how exposure intensity is generated. Figure 1 shows representative scenes; the subsections summarize phobia-specific hierarchy, interaction logic, and logged variables.



Figure 1. Representative scenes from the three prototypes: (a) cat phobia, (b) height phobia, and (c) flight phobia.

1. Cat-Phobia Prototype

The cat system progresses from pictures and a distant, passive cat to closer, moving, interactive exposure. Visibility, distance, orientation, motion, and behavior define reproducible stages; the key challenge is balancing engaging agent behavior with predictability. Logs record stage time, completion or interruption, cat distance, approach or withdrawal, and interaction completion, directly reflecting the object-based hierarchy.

2. Height-Phobia Prototype

The height system progresses from protected to increasingly exposed elevated scenes. Intensity combines elevation, visual depth, openness, edge proximity, barriers, ground visibility, and movement, while stable navigation limits motion-related discomfort. Users can stop or step back. Logs record edge distance, time at height, entry into exposed areas, retreat, and level progression, reflecting the spatial structure.

3. Flight-Phobia Prototype

The flight system models flying as a timed sequence of cabin exposure, announcements, taxiing, takeoff, in-flight sounds, and turbulence. Its main challenge is synchronizing visual, auditory, and temporal cues. Users may pause, repeat, reduce intensity, or exit without controlling the aircraft. Logs record stage duration and completion, pauses, repetitions, interruptions, termination point, and responses to takeoff or turbulence.

V. RESULTS: CROSS-CASE ANALYSIS AND PRELIMINARY DESIGN FRAMEWORK

1. Shared and Phobia-Specific Logic

Across systems, the shared design core includes graded progression, repeatable stages, predictable transitions, participation control, stopping mechanisms, and general event logging. These functions are content-independent, whereas the mechanism generating exposure intensity remains stimulus-specific. Object-based exposure varies with a feared agent; spatial exposure depends on the user's position and perception; event-sequence exposure depends on timing across phases. This structure determines hierarchy, parameters, interaction design, safety risks, and meaningful behavioral measures.

Table 1. Cross-case comparison of exposure structures and design requirements.

Dimension	***Cat phobia***	***Height phobia***	***Flight phobia***
Exposure structure	Dynamic object	Spatial environment	Temporal event sequence
Exposure unit	Cat proximity and behavior	Position within an elevated scene	Phase of the simulated flight
Main parameters	Visibility, distance, movement, orientation, behavior	Elevation, depth, openness, edge distance, barriers	Stage, sound, motion cues, turbulence
Primary challenge	Believable but predictable agent behavior	Stable navigation and calibrated spatial intensity	Timing and multimodal synchronization
Implemented VR logging	Distance, approach, withdrawal, interaction completion	Edge distance, exposed-area entry, retreat, duration	Completion, pauses, interruptions, termination point
Engineering category	Object-based	Spatial	Event-sequence

Commonly implemented functions across all systems include synchronized VR/wearable logs, session timestamps, persistent backend storage, and access to session history.

2. Proposed Framework

The comparison informed a five-stage development framework:

- Characterize the feared experience as primarily object-based, spatial, event-sequence, or hybrid.
- Construct an exposure hierarchy with stages representing meaningful increases in challenge or uncertainty.
- Map stages to explicit controllable parameters where possible, supporting reuse, testing, and adaptation.

- Define interaction, progression, and safety rules, including how users enter, repeat, pause, or leave a stage, and distinguishing control over participation from control over the feared event.
- Define common session events and additional behavioral measures that reflect the relevant exposure structure.

The second-generation implementation extends the framework into three layers: a stimulus-specific VR layer for object-based, spatial, or event-sequence logic; a session and synchronization layer for VR events and wearable data; and a backend layer for synchronized records and session histories. A recommendation component is under development (Figure 2).

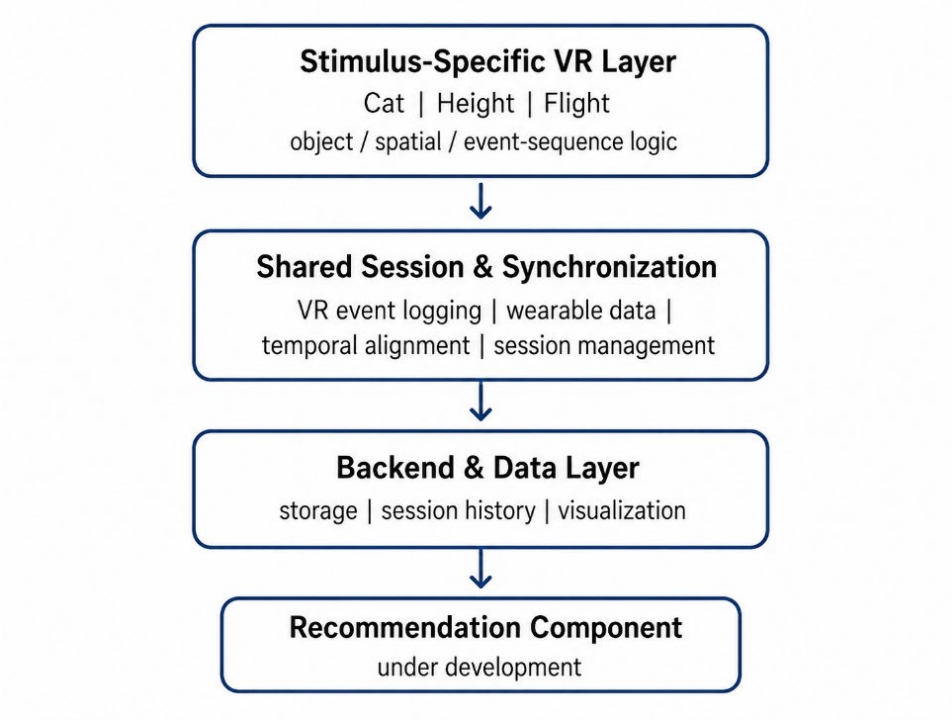


Figure 2. Three-layer architecture of the second-generation VR exposure systems

3. Derived Guidelines and Scope

Three guidelines follow: model exposure stages through parameters and transition rules rather than only separate scenes; calibrate realism to the exposure objective while avoiding unpredictable behavior, unnecessary complexity, and cybersickness; and plan logging with the hierarchy so recorded behavior reflects approach, avoidance, persistence, or interruption. The categories may overlap, and the framework remains an initial engineering model rather than a clinical protocol; decisions about progression require expert review and empirical validation.

VI. DISCUSSION

The results indicate that VR phobia systems differ mainly in the mechanism generating intensity, not simply in visual content. The second-generation implementation separates phobia-specific exposure logic from common

session, synchronization, and backend functions, enabling modular development without forcing different fears into a single exposure structure. This is preliminary technical support, not usability or clinical validation. Parameterizing stages may improve consistency, testing, and later personalization by making changes in distance, openness, edge proximity, sound, or turbulence explicit and reproducible; adaptive progression rules still require empirical validation. Two HCI implications follow. Realism should support recognition and engagement without adding unpredictability or discomfort, because greater immersion does not necessarily improve clinical outcomes [5],[8]. User control should preserve participation and safety without eliminating the feared event; for example, flight exposure can be paused or stopped while the aircraft remains under external control. Behavioral logging is most useful when aligned with exposure structure: common events enable cross-system comparison, while distance, edge proximity, retreat, and interruption capture phobia-specific behavior. Synchronizing these traces with wearable data and preserving session histories provides a basis for future multimodal analysis and recommendation logic. Still, interpretation must be evaluated with usability, self-report, and clinical input.

VII. LIMITATIONS AND FUTURE WORK

The framework is derived from only three heterogeneous student-developed prototypes, which do not represent all specific phobias or hybrid cases and were not evaluated with common technical benchmarks. The qualitative analysis relied on artifacts and documentation; the cat developer's five-session experience was a single, non-independent formative observation. No controlled participant study assessed usability, presence, cybersickness, fear activation, behavioral change, or clinical outcomes. The second-generation logging, wearable integration, synchronization, backend histories, and data interpretation also require systematic evaluation, including synchronization accuracy, missing-data handling, repeated-session reliability, backend usability, and user acceptance. Future work includes a clinical expert review, followed by the planned usability study with predefined measures. The recommendation mechanism remains under development and will require separate evaluation. The architecture should also be tested across additional phobias and hybrid scenarios before any formal clinical efficacy study.

VIII. CONCLUSION

This work identifies object-based, spatial, and event-sequence exposure structures and derives a preliminary framework separating common session and data functions from phobia-specific intensity and interaction logic. Second-generation implementations add synchronized VR/wearable logging, backend storage, and session histories, providing initial technical support for this separation. The contribution remains engineering-oriented work in progress; expert review, usability and feasibility testing, data-pipeline evaluation, and clinical assessment are still required.

REFERENCES

[1]. Odgers, K., Kershaw, K.A., Li, S.H., Graham, B.M.: The relative efficacy and efficiency of single- and multi-session exposure therapies for specific phobia: A meta-analysis. Behavior Research and Therapy 159, 104203 (2022). https://doi.org/10.1016/j.brat.2022.104203

[2]. Craske, M.G., Treanor, M., Zbozinek, T.D., Vervliet, B.: Optimizing exposure therapy with an inhibitory retrieval approach and the OptEx Nexus. Behavior Research and Therapy 152, 104069 (2022). https://doi.org/10.1016/j.brat.2022.104069

[3]. Wiebe, A., Kannen, K., Selaskowski, B., et al.: Virtual reality in the diagnosis and therapy for mental disorders: A systematic review. Clinical Psychology Review 98, 102213 (2022). https://doi.org/10.1016/j.cpr.2022.102213

[4]. Schröder, D., Wrona, K.J., Müller, F., Heinemann, S., Fischer, F., Dockweiler, C.: Impact of virtual reality applications in the treatment of anxiety disorders: A systematic review and meta-analysis of randomized-controlled trials. Journal of Behavior Therapy and Experimental Psychiatry 81, 101893 (2023). https://doi.org/10.1016/j.jbtep.2023.101893

[5]. Kuleli, D., Tyson, P., Davies, N.H., Zeng, B.: Examining the comparative effectiveness of virtual reality and in-vivo exposure therapy on social anxiety and specific phobia: A systematic review and meta-analysis. Journal of Behavioral and Cognitive Therapy 35, 100524 (2025). https://doi.org/10.1016/j.jbct.2025.100524

[6]. Lacey, C., Frampton, C., Beaglehole, B.: oVRcome – Self-guided virtual reality for specific phobias: A randomized controlled trial. Australian & New Zealand Journal of Psychiatry 57(5), 736-744 (2023). https://doi.org/10.1177/00048674221110779

[7]. Francová, A., Kolman Jablonská, M., Lhotská, L., Husák, J., Fajnerová, I.: Efficacy of exposure scenario in virtual reality for the treatment of acrophobia: A randomized controlled trial. Journal of Behavior Therapy and Experimental Psychiatry 88, 102035 (2025). https://doi.org/10.1016/j.jbtep.2025.102035

[8]. Mesquita, C., Ventura, P., da Silva, H.C., et al.: Is virtual reality always better than 2D videos? A head-to-head randomized controlled trial for fear of flying with long-term follow-up. Psychiatry Research 354, 116784 (2025). https://doi.org/10.1016/j.psychres.2025.116784

[9]. Ebrahimi, A., Akbarzadeh, F., Asgharipour, N., Salehabadi, R., Arjamandi, M.: Investigating the effectiveness of ailurophobia treatment using virtual reality technique compared to metacognitive therapy: A randomized clinical trial. BMC Psychology 13, 50 (2025). https://doi.org/10.1186/s40359-025-02378-9

[10]. Arnfred, B., Svendsen, J.K., Adjourlu, A., Horthøj, C.: Scoping review of the hardware and software features of virtual reality exposure therapy for social anxiety disorder, agoraphobia, and specific phobia. Frontiers in Virtual Reality 4, 952741 (2023). https://doi.org/10.3389/frvir.2023.952741

[11]. Graham, W.M., Drinkwater, R., Kelson, J., Kabir, M.A.: Self-guided virtual reality therapy for anxiety: A systematic review. International Journal of Medical Informatics 200, 105902 (2025). https://doi.org/10.1016/j.ijmedinf.2025.105902

[12]. Simón-Vicente, L., Rodríguez-Cano, S., Delgado-Benito, V., Ausín-Villaverde, V., Cubo Delgado, E.: Cybersickness: A systematic literature review of adverse effects related to virtual reality. Neurología (English Edition) 39(8), 701-709 (2024). https://doi.org/10.1016/j.nrleng.2022.04.007

[13]. Guillen-Sanz, H., Checa, D., Miguel-Alonso, I., Bustillo, A.: A systematic review of wearable biosensor usage in immersive virtual reality experiences. Virtual Reality 28, 74 (2024). https://doi.org/10.1007/s10055-024-00970-9

[14]. Kazemi, R., Lee, S.C.: Human Factors/Ergonomics (HFE) Evaluation in the Virtual Reality Environment: A Systematic Review. International Journal of Human–Computer Interaction 40, 4533–4549 (2024). https://doi.org/10.1080/10447318.2023.2227835

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Semantic Interoperability of Heterogeneous Research Data: A FAIR-Aligned Ontology-Driven Framework for Interdisciplinary Research

Kevin Aber
Department of Computer Science, Faculty ot Science, Gulu University,
Gulu, Uganda
k.aber@gu.ac.ug 0009-0000-9567-1672

Lone Dirckinck Holmfeld
Department of Culture and Communication | Faculty of the Social Science and Humanities Aalborg University,
Copenhagen, Denmark
lone@ikk.aau.dk

Geoffrey Olok Tabo
Department of Computer Science, Faculty ot Science, Gulu University,
Gulu, Uganda
g.tabo@gu.ac.ug

Christine Kalumera Akello
Directorate of Information and Communication Technology, Gulu University
Gulu, Uganda
c.k.akello@gu.ac.ug

Benedict Oyo
Department of Computer Science, Faculty ot Science, Gulu University,
Gulu, Uganda
b.oyo@gu.ac.ug

Daniel Ogenrwot
Department of Computer Science, Faculty of Science, Gulu University
Gulu, Uganda
d.ogenrwot@gu.ac.ug

Abstract
Interdisciplinary research produces heterogeneous datasets on independently managed platforms, creating semantic barriers to effective integration, reuse, and knowledge discovery. Current ontology-based integrative approaches have shown good performance in structured and domain knowledge environments, but have limitations in integrating qualitative, quantitative, and contextual research knowledge. This study presents the Green Charcoal Ontology (GCO), a FAIR-aligned ontology-driven semantic mediation framework developed using an interdisciplinary Green Charcoal project as an empirical case. The framework incorporates production, market, and geospatial research concepts using a modular OWL 2 DL ontology that links locally developed research vocabularies to existing Semantic Web standards such as SOSA, PROV-O, and GeoSPARQL. The development of the ontology was based on the Design Science Research Methodology and the NeOn approach, and its evaluation was conducted using OWL reasoning, SHACL-based validation, and competency-question-driven SPARQL queries. Results show that cross-domain knowledge retrieval from the integrated knowledge graph is successful, that the data constraints are evaluated, and that logical consistency is achieved across 2674 axioms. In addition to semantic integration, GCO offers a conceptual design for Research Data Management Systems, including common concepts, relationships, provenance structures, and validation mechanisms. The study shows that FAIR-compatible Semantic Interoperability can be achieved through practice-informed ontology engineering while maintaining domain-specific knowledge practices in interdisciplinary and resource-constrained research scenarios.

Keywords: Semantic Interoperability, Ontology Engineering, FAIR Data, principles, Knowledge Graph, Interdisciplinary Research, Research Data Management System

I. INTRODUCTION

The global research landscape is increasingly interdisciplinary, bringing together diverse scientific communities that generate heterogeneous data across institutional repositories, sensor networks, spreadsheets, and relational databases. While this diversity accelerates scientific discovery, it also introduces semantic heterogeneity that hinders data integration, interoperability, and reuse. Beyond a technical issue, semantic heterogeneity reflects differences in disciplinary practices, conceptualizations, and interpretations. Effective interdisciplinary collaboration, therefore, requires shared mechanisms that facilitate communication without enforcing complete conceptual uniformity. The concept of boundary objects explains how shared representations enable collaboration across communities while preserving domain-specific perspectives. Ontology-based semantic models can serve as formal boundary objects by providing machine-interpretable conceptual frameworks that support semantic interoperability and knowledge exchange across disciplines [1]. Nevertheless, achieving semantic interoperability remains a major challenge in interdisciplinary research environments. Unlike syntactic heterogeneity, which arises from differences in data formats and encodings, semantic heterogeneity occurs when the same concepts are represented with different meanings or incompatible conceptual structures [2]. Conventional data integration approaches, such as extract-transform-load (ETL) pipelines, primarily address structural and syntactic inconsistencies but provide limited support for resolving semantic conflicts [3]. Consequently, datasets may be technically integrated while remaining semantically incompatible. Recent studies identify semantic interoperability as a persistent challenge and highlight ontology mapping as an effective mechanism for aligning heterogeneous knowledge representations [4], [5]. Ontologies also function as socio-technical boundary objects that facilitate shared understanding among research communities while preserving disciplinary perspectives [6], [7]. Existing approaches, including hybrid ontology-based integration frameworks [8] and ontology database mapping techniques [9], have advanced semantic integration. However, they primarily target structured or domain-specific environments and provide limited support for interdisciplinary research settings that integrate qualitative, quantitative, and geospatial data while preserving contextual meaning. The FAIR (Findable, Accessible, Interoperable, and Reusable) principles provide internationally accepted guidelines for improving the quality, accessibility, and reusability of research data [10]. However, FAIR compliance extends beyond metadata annotation and open access to include semantic mechanisms that preserve meaning across heterogeneous datasets [11]. Semantic Web standards, including the Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL), provide the foundation for machine-actionable knowledge representation and semantic interoperability. Despite these advances, few ontology-driven FAIR frameworks explicitly address semantic mediation across interdisciplinary and resource-constrained research environments, particularly within African research institutions. This challenge is exemplified by the Unlocking the Potential of Green Charcoal to Mitigate Climate Change in Northern Uganda (UPCHAIN) project, an interdisciplinary initiative comprising three complementary work packages. WP1 focuses on green charcoal production and quality assessment; WP3 investigates market dynamics and livelihood impacts; and WP5 analyses raw material availability using Participatory Geographical Information Systems (PGIS). Although these work packages contribute to a common research objective, they employ distinct terminologies, methodologies, metadata structures, and conceptual models. This diversity creates semantic heterogeneity that limits cross-domain data integration and knowledge reuse while providing an ideal empirical setting for developing and validating a unified semantic framework. Although ontology-based integration has demonstrated success in enterprise systems and well-defined knowledge domains, comparatively little attention has been paid to interdisciplinary research environments in which heterogeneous observational, socio-economic, experimental, and geospatial data must be integrated without compromising domain-specific semantics. Existing studies rarely combine ontology engineering, FAIR principles, semantic validation, and competency-question-driven evaluation within a unified semantic interoperability framework. This study addresses this gap by proposing an ontology-driven semantic mediation framework for heterogeneous research data environments using the UPCHAIN project as a real-world case study.

The objectives of this study are to: (i) analyze semantic heterogeneity across interdisciplinary work packages; (ii) develop a FAIR-aligned Green Charcoal ontology; (iii) validate the ontology using OWL reasoning and SHACL constraint checking; and (iv) evaluate semantic interoperability through competency-question-driven

SPARQL queries. The study contributes to the field by first developing a modular ontology that integrates the production, market, and geospatial research domains while reusing established Semantic Web vocabularies, including SOSA, PROV-O, and GeoSPARQL. Second, it proposes an ontology-driven semantic interoperability framework that supports FAIR-compliant integration of heterogeneous interdisciplinary research data. Third, it demonstrates a comprehensive validation workflow combining OWL reasoning, SHACL constraint validation, and competency-question-based SPARQL queries to verify logical consistency and semantic correctness. Finally, it provides a practice-informed case study showing how ontology-driven semantic mediation can improve interoperability and knowledge sharing in interdisciplinary and resource-constrained research environments.

II. RELATED WORK

1. Semantic Interoperability in Interdisciplinary Research

Modern scientific collaborations generate diverse datasets that are stored in independently managed repositories using different data models, metadata standards, measurement conventions, and disciplinary terminologies [12]. This diversity creates significant conceptual fragmentation, particularly in the life sciences, where repositories such as BioPortal contain numerous independently developed ontologies [13], [14]. Similar fragmentation exists in scholarly infrastructures such as the Microsoft Academic Graph [15]. It is even more pronounced in the social sciences and humanities, where concepts are often context-dependent and interpreted differently across disciplines. Interoperability is commonly classified into technical, syntactic, semantic, organizational, and conceptual levels [16]. Although standards such as XML and JSON-LD effectively support data exchange, they cannot guarantee that information retains the same meaning across systems [17]. Consequently, successful collaboration depends on sharing knowledge rather than merely exchanging data [18]. Traditional integration approaches, including relational databases and extract-transform-load (ETL) pipelines, resolve structural inconsistencies but rarely preserve contextual meaning or conceptual relationships among datasets [19]. Semantic technologies address this limitation by representing concepts and their relationships in a machine-interpretable form. Beyond their technical role, ontologies also function as boundary objects, providing shared conceptual models that facilitate collaboration while allowing individual disciplines to retain their own perspectives [20]. Recent advances in Semantic Web technologies have improved knowledge integration across distributed information systems. Nevertheless, most existing frameworks focus on enterprise applications or single-domain repositories. Limited attention has been paid to collaborative environments that integrate independently managed datasets, provenance requirements, governance mechanisms, and FAIR principles into a unified semantic framework. Addressing these challenges requires ontology-driven approaches that support meaningful knowledge integration, cross-domain discovery, and sustainable data reuse. This study responds to this need by developing a FAIR-aligned semantic integration framework for heterogeneous research data.

2. Ontology-Based Data Integration

Ontology-Based Data Integration (OBDI) addresses conceptual inconsistencies by introducing a shared semantic layer that provides a unified view of heterogeneous information sources. Two principal integration paradigms have emerged. Ontology-Based Data Access (OBDA) enables virtual querying of distributed databases through declarative mappings and knowledge graphs. In contrast, materialized integration transforms heterogeneous datasets into a common RDF repository or data warehouse for centralized querying [8], [21], [22]. Both approaches allow users to formulate semantic queries independently of the underlying storage technologies. A typical OBDI architecture comprises a domain ontology, commonly represented in OWL 2 QL. These declarative mappings relate ontology concepts to source schemas, and heterogeneous repositories such as relational, JSON, and graph databases [23], [24]. Depending on the integration strategy, architectures may adopt global, local, or hybrid ontology models. Global ontologies simplify query processing through a unified

conceptual schema, while local ontologies preserve domain autonomy. Hybrid architecture combines these advantages by linking independent domain ontologies through a shared upper-level model, thereby improving modularity, scalability, and semantic consistency across distributed repositories [25]. OBDI has been widely applied in enterprise information systems, healthcare, and industrial applications, demonstrating its effectiveness in semantic querying and knowledge integration. However, its adoption in interdisciplinary research data management remains limited. Existing approaches primarily emphasize data access and query processing, with comparatively little attention given to FAIR-compliant governance, provenance management, modular ontology engineering, and comprehensive semantic validation. These limitations highlight the need for integration frameworks that not only enable semantic access to heterogeneous datasets but also support trustworthy, reusable, and interoperable research data throughout its lifecycle.

3. Ontology Engineering Methodologies and Semantic Web Standards

Ontology engineering methodologies provide systematic guidance for developing interoperable semantic models throughout their lifecycle. Early approaches, including Methontology [26] and Ontology 101 [27], established structured processes covering ontology specification, conceptualization, formalization, implementation, and evaluation. Although these methodologies remain influential, they were primarily designed for standalone ontology development and provide limited support for modular reuse and collaborative development across distributed domains. The NeOn methodology [28] extends these approaches by promoting ontology reuse, modularization, and collaborative engineering, making it particularly suitable for interdisciplinary environments comprising multiple autonomous knowledge domains. Consequently, NeOn provides the methodological foundation for the work-package-oriented ontology developed in this study. Modern ontology engineering is underpinned by the W3C Semantic Web standards, which collectively support interoperable knowledge representation. RDF provides graph-based data modeling [29], OWL enables formal knowledge representation and logical reasoning [30], while SPARQL facilitates semantic querying. Complementary standards, including SKOS, SHACL, PROV-O, SOSA, and GeoSPARQL, support vocabulary management, constraint validation, provenance tracking, sensor observations, and spatial representation, respectively. In addition, declarative mapping languages such as R2RML and RML enable heterogeneous data sources to be transformed into interoperable RDF knowledge graphs, facilitating scalable semantic integration workflows [9]. Recent developments increasingly emphasize modular ontology design and the reuse of established vocabularies to improve maintainability, interoperability, and long-term sustainability. However, many existing implementations focus primarily on ontology construction and knowledge representation, with less emphasis on integrating ontology engineering, semantic validation, provenance management, and FAIR principles within a unified framework for interdisciplinary research. Addressing this limitation requires methodologies that support not only ontology development but also semantic interoperability throughout the research data lifecycle.

4. FAIR Research Data Management

The FAIR principles provide a framework for improving the Findability, Accessibility, Interoperability, and Reusability of research data through persistent identifiers, rich metadata, and machine-actionable standards [10]. Increasingly, FAIR is viewed not merely as a data management guideline but as a socio-technical framework that relies on shared conceptual models to enable meaningful data exchange and reuse across diverse research communities [31], [32]. Its implementation is typically supported by FAIRification workflows that involve data curation, metadata enrichment, semantic annotation, and ontology-based integration. Evaluation frameworks, including the RDA FAIR Data Maturity Model and GO FAIR guidelines, together with assessment tools such as FAIR-Checker [33], FAIRshake, F-UJI, and O'FAIRe, provide measurable indicators of compliance using technologies such as SPARQL and SHACL. More recent developments, including FAIR Digital Objects and FAIR Data Points, further strengthen machine-actionability and support federated data ecosystems [34]. Despite these advances, interoperability remains the least mature FAIR dimension because aligning data meaning across heterogeneous domains is inherently challenging. Most implementations achieve metadata harmonization but

provide limited support for semantic integration across independently developed knowledge models. These limitations are particularly evident in interdisciplinary and resource-constrained research environments, where differences in conceptual models, infrastructure, and technical expertise hinder effective knowledge sharing [35], [36]. Consequently, ontology-driven semantic frameworks remain essential for translating FAIR principles into meaningful, machine-interpretable interoperability that supports cross-domain integration and long-term data reuse.

5. Existing Ontology-Based Research Data Management Systems

Ontology-based research data management (RDM) systems have evolved from standalone ontology development tools to integrated FAIR-enabled knowledge graph platforms. Early systems, including Protégé, SOR, QuOnto, and Oracle Semantic Technologies, established the foundations for ontology development, reasoning, and semantic data management. Community-driven repositories such as the OBO Foundry further promoted ontology reuse, standardization, and collaborative development. More recently, ontology-based data access and management platforms, including Ontop and Mastro, have enabled scalable semantic access to heterogeneous relational databases through declarative mappings, supporting applications in domains such as energy, transportation, and manufacturing [37]. Complementary technologies, including RML, RDFAdaptor, and SHACL-based validation pipelines, further facilitate heterogeneous data transformation and quality assurance [38]. Recent advances have extended these capabilities by integrating FAIR principles with knowledge graphs and research data infrastructures. Platforms such as LDM, NFDI4Earth, PO²/TransformON, and ARGOS support FAIR-compliant data management through semantic annotation, metadata standardization, and lifecycle management, thereby improving data discovery, sharing, and reuse [32], [39]. Collectively, these systems demonstrate substantial progress towards machine-actionable and interoperable research data ecosystems. Despite these advances, several limitations remain. Many existing platforms prioritize data publication and semantic access while providing limited support for integrated ontology engineering, modular knowledge organization, semantic validation through OWL reasoning and SHACL constraints, and competency-question-driven evaluation within a single framework. In addition, relatively few implementations have been evaluated in interdisciplinary research settings characterized by independently managed datasets, diverse conceptual models, and resource-constrained institutional environments. These limitations motivate the development of ontology-driven frameworks that combine FAIR principles, semantic interoperability, governance, and comprehensive validation to support collaborative research data management across multiple domains.

III. METHODOLOGY

1. Research Design

This study adopts the Design Science Research Methodology (DSRM) (Hevner, 2007) as the overarching research framework and the NeOn ontology engineering methodology (Nasir et al., 2024; Suárez-Figueroa et al., 2015) to guide the development of the Green Charcoal Ontology (GCO). DSRM provides the overall process for designing, implementing, demonstrating, evaluating, and communicating the semantic interoperability artifact, while NeOn provides ontology-specific guidance for conceptualization, reuse, modularization, and implementation. Thus, NeOn complements DSRM by operationalizing the ontology engineering activities within the DSRM design and development phase. Following DSRM, the research process comprises problem identification, objective definition, design and development, demonstration, evaluation, and communication. These stages represent a context-specific operationalization of NeOn principles for transforming heterogeneous green charcoal research knowledge into a formal semantic artifact. After development, the GCO was demonstrated and evaluated using UPCHAIN research scenarios through logical consistency checking, constraint validation, and competency questions. The findings and artifacts were subsequently documented and communicated as a contribution toward FAIR-aligned interdisciplinary research data management.

2. Problem Identification

The problem identification phase explored research data management and semantic interoperability issues in the UPCHAIN project. Five focus group discussions (FGDs) with researchers from WP1, WP3, and WP5 were held, and existing work package data structures were explored via entity–relationship (ER) diagrams. The FGDs discussed domain concepts, data generation processes, terminology, and interactions between work packages. Thematic analysis and ER mapping were used to systematically understand the research domain and the need for a semantic interoperability artifact.

3. Objective Definition

The goal of this research is to design and develop an ontology-driven semantic model that facilitates interoperability, integration of heterogeneous Green Charcoal research data, and meaningful reuse across WP1, WP3, and WP5. The study aims to build a common semantic basis by creating a modular OWL 2 DL ontology network to define concepts, relationships, provenance, measurements, and quality constraints in the domain. The developed ontology is tested through reasoning, SHACL validation, and competency questions to assess its ability to support the consistent representation of data and the integration of knowledge across domains.

4. Design and Development

The NeOn ontology engineering methodology was used during this DSRM phase to guide the construction of the GCO. DSRM introduced the overall research framework, but NeOn was responsible for implementing the ontology based on four adapted stages as follows: (1) requirements analysis; (2) conceptualization and design; (3) implementation using OWL 2 DL. This method guaranteed the systematic development of the ontology, starting from stakeholder requirements to a validated semantic artifact.

5. Ontology Engineering

Ontology requirements were established through thematic analysis of five focus group discussions (WP1: two FGDs with seven participants; WP5: two FGDs with six participants; and WP3: one FGD with five participants) and by mapping existing work package data structures using entity–relationship diagrams. The iterative coding analysis identified recurring themes, key domain concepts, relationships, and variations in terminology, while ER mapping guided ontology class identification and integration points. Four dimensions of data heterogeneity: structural (variations in data organization), semantic (divergent interpretations of shared concepts), syntactic (discrepancies in data formats and encoding), and provenance (inconsistencies in data lineage representations), shaped a modular ontology that enhances interoperability while preserving domain-specific integrity. The ontology design incorporated established vocabularies, including SOSA/SSN for observations, PROV-O for provenance, and GeoSPARQL for spatial representation. It was implemented using OWL 2 DL with defined classes, properties, and constraints to support semantic interoperability across heterogeneous research datasets. The resulting ontology harmonizes datasets from WPs into a unified schema, ensuring traceability and facilitating spatial analysis for sustainability assessment.

6. Ontology Design

Ontology Specification

The ontology specification phase defines the ontology's purpose, scope, objectives, intended use, and functional requirements (Nie et al., 2024). It establishes the domain boundaries and translates the identified data

heterogeneity challenges across the work packages into competency questions, stakeholder requirements, and modeling objectives that guide ontology development. The GCO focuses on interdisciplinary research on green charcoal, integrating environmental, agricultural, engineering, and socio-economic knowledge across the value chain, from biomass sourcing to market distribution. Its primary objective is to formulate the competency questions (CQs) developed by the researchers during the ontology requirements analysis phase to define the scope and knowledge-retrieval capabilities of the GCO. They were derived from UPCHAIN research workflows, focus group discussions (FGDs), analysis of existing datasets, and interactions with researchers across work packages, which identified the key knowledge relationships required to integrate production, market, and spatial information. To test the ability of the GCO to facilitate semantic integration across work-packages, two representative CQs were chosen for this study:

- CQ1: What is the complete provenance chain of a green charcoal product, including biomass feedstock species and sourcing location, carbonization process and responsible agents, production sites, quality monitoring records, and final market distribution to end users?
- CQ2: How does green charcoal product quality, together with the sustainability characteristics of biomass feedstock, influence market demand, pricing, and end-user adoption across different market segments in Northern Uganda?

The CQs supported the ontology scope definition process by determining which concepts, properties, relationships, and constraints were necessary. They were then used as assessment criteria to evaluate the ontology's ability to support interdisciplinary knowledge retrieval. Independent domain experts will refine all these during the Research Data Management System implementation.

Ontology Design

The GCO was designed as a modular semantic model to represent and integrate knowledge across WPs, operating as independent modules linked through defined relationships while maintaining domain autonomy. The ontology consists of three interconnected modules: WP1 (Production and Quality Control), WP3 (Markets and Livelihoods), and WP5 (Geospatial and Raw Materials sharing a common semantic foundation, with hierarchical classes representing general concepts (abstract superclasses), domain-specific subclasses, and specialized leaf classes such as GCProductionProcess, GCMarketActor, and BiomassZone class for spatial reasoning in line with GeoSPARQL geo: Feature. Cross-module object properties: suppliesReceipt, resultsInReceipt, locatedInZone, correspondsToReceipt, and tracksSalesOf were defined to connect production, market, and geospatial knowledge, enabling semantic integration while preserving modularity, extensibility, and clear ownership of domain knowledge. Figure 1 illustrates the logical architecture of the Green Charcoal Ontology, showcasing the three domain modules, their internal class hierarchy, and the semantic relationships linking them, with a focus on conceptual organization rather than software implementation.

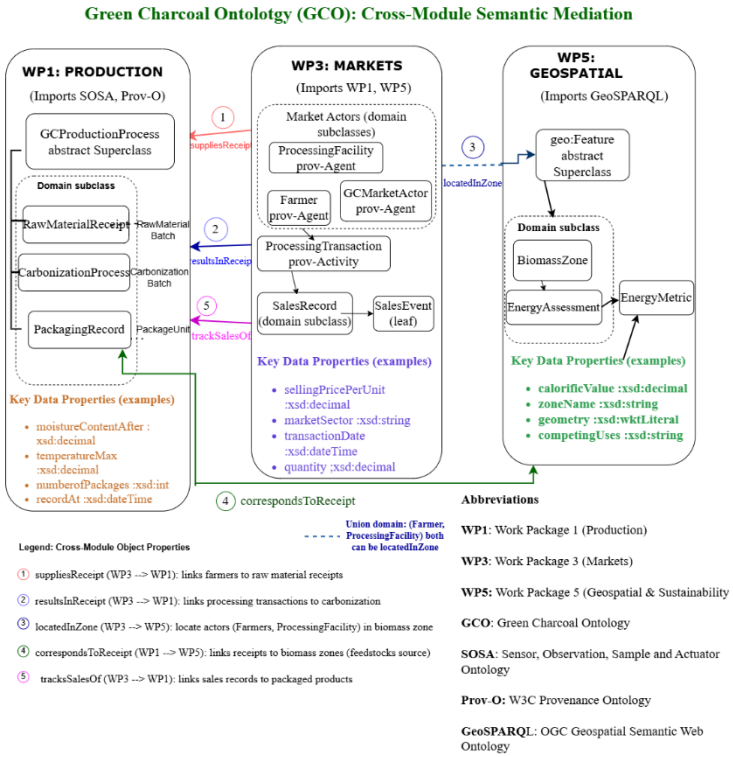


Figure 1. Green Charcoal Ontology (GCO) modular design architecture showing domain modules and cross-module semantic relationships. The logical design specifies the semantic interactions required to integrate interdisciplinary knowledge. Importantly, the union-based domain of locatedInZone allows multiple types of entities, such as farmers and processing facilities, to be associated with a shared spatial location, thereby facilitating spatial inference in production and market activities. Likewise, suppliesReceipt relates market actors to biomass receipts for tracking and reporting purposes; resultsInReceipt associates various production processes with their products; correspondsToReceipt links spatial information about biomass to sourcing information; and tracksSalesOf links market transactions to the produced charcoal products. Together, these semantic relations enable knowledge sharing in the three domains while maintaining the independence of the ontology modules. The ontology design is also based on existing Semantic Web vocabularies to maximize interoperability and reduce duplication of semantic concepts. For observations and sensing activities, SOSA/SSN was selected, PROV-O for the representation of provenance, and GeoSPARQL for spatial entities and spatial relationships. The use of these common vocabularies provides compatibility with existing semantic resources and consistency with FAIR-oriented ontology engineering practices.

7. Ontology Implementation

The GCO was implemented using the W3C Web Ontology Language (OWL) framework, specifically OWL 2 DL, for formal semantic representation, decidable reasoning support, and interoperability with Semantic Web technologies. The ontology was developed in Protégé v5.6.9, which provided the environment for defining ontology classes, object properties, data properties, class hierarchies, and logical axioms. It also supported ontology visualization, ontology imports, consistency checking, and iterative refinement throughout the implementation process. The modules were implemented as separate OWL 2 DL files, serialized in Turtle (TTL), to support portability, reuse, and integration with Semantic Web environments. The modules maintained independent domain ownership while enabling semantic integration through controlled ontology imports and

cross-module object properties. The class hierarchy follows a three-tier structure of superclasses, domain subclasses, and leaf specializations. At the same time, object and datatype properties are defined using XML Schema datatypes *(xsd: decimal, xsd: string, xsd:dateTime)*. Cross-module semantic interactions were implemented through five object properties, the only permitted relationships to span ontology modules. All other object properties remained within their respective modules, preserving modularity and minimizing semantic coupling. Restricting inter-module interaction to these five semantic bridges reduced reasoning complexity, simplified ontology maintenance, and enabled cross-module competency-question evaluation using standard SPARQL queries. Logical constraints identified during ontology design were encoded using OWL 2 DL class restrictions, enabling automatic reasoning over ontology concepts. To complement OWL's open-world semantics, data integrity constraints were implemented separately using SHACL node shapes, enabling validation of instance data against predefined business rules, such as quality thresholds and mandatory property constraints. Specifically, WP3 imports WP1 and WP5 to access shared concepts and to establish relationships among the production, market, and geospatial knowledge domains, while preserving the separation of concerns. Established Semantic Web vocabularies were imported into the GCO. SOSA represents production observations and quality assessments, sensing activities, and measurements, PROV-O captures provenance across production, market, and geospatial processes through Activity, Agent, and Entity to support traceability of research data, and GeoSPARQL supports spatial modeling of BiomassZone for location-based reasoning. These reused vocabularies strengthened alignment with existing semantic standards and supported FAIR representation. Figure 2 (a-c) presents the implemented GCO ontology graph, illustrating the internal structure of the three modules, imported vocabularies, and the semantic connections established between them. WP1 module in Figure 2a shows the production pipeline with eight distinct stages (receipt, milling, mixing, binder preparation, extrusion, drying, carbonization, transport), each generating its own data. Therefore, GCProductionProcess is an abstract superclass (the green-highlighted node in Figure 2a), with each stage as a subclass, and can be queried as a whole or stage-by-stage. The Activity node in Figure 2a shows that each subclass also inherits prov: Activity, since every stage is, in provenance terms, an activity that consumes one entity and produces another. The chain can be traversed using standard PROV-O properties (wasGeneratedBy, wasInformedBy) instead of custom WP1-only properties. Figure 2a also shows the populated data consisting of six CarbonizationProcess and six RawMaterialReceipt individuals, one per production chain.

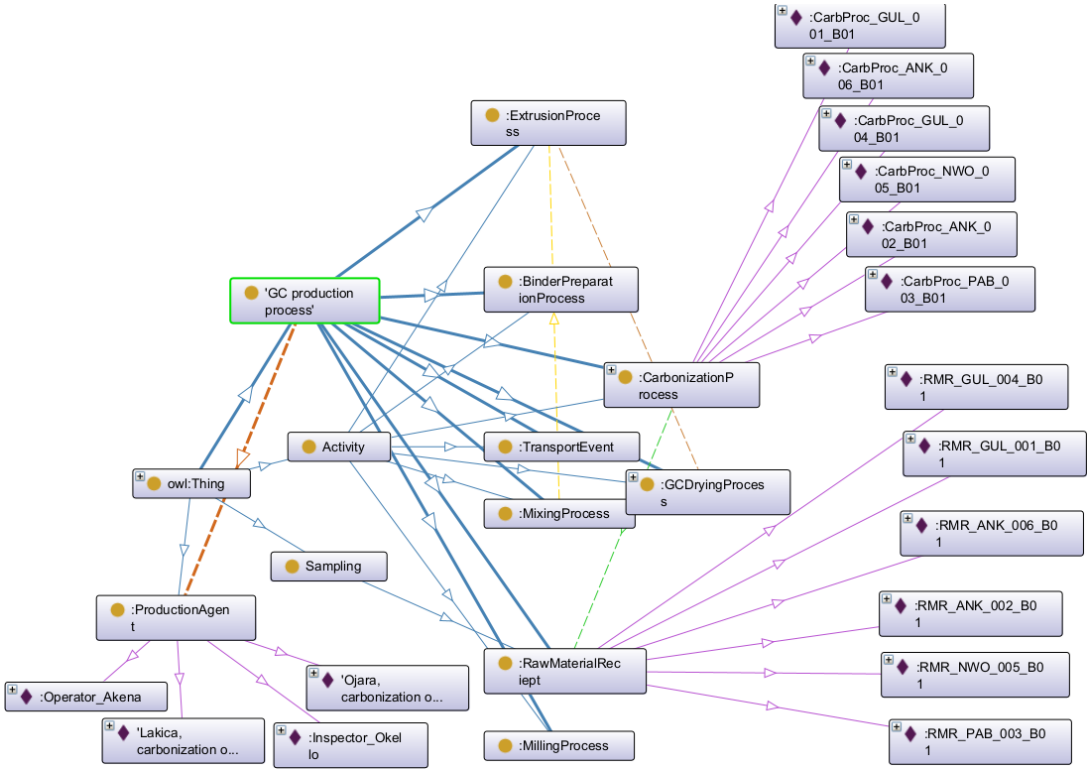


Figure 2a. WP1 Production Process Hierarchy and Populated Carbonization Process and Raw Material Receipt Instances

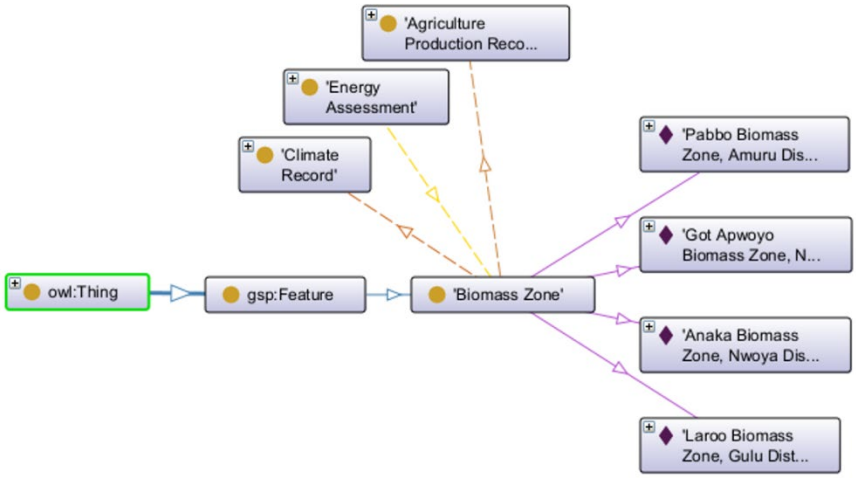


Figure. 2b. WP5 BiomassZone as a subclass of GeoSPARQL geo: Feature, with aggregated climate, production, and energy records

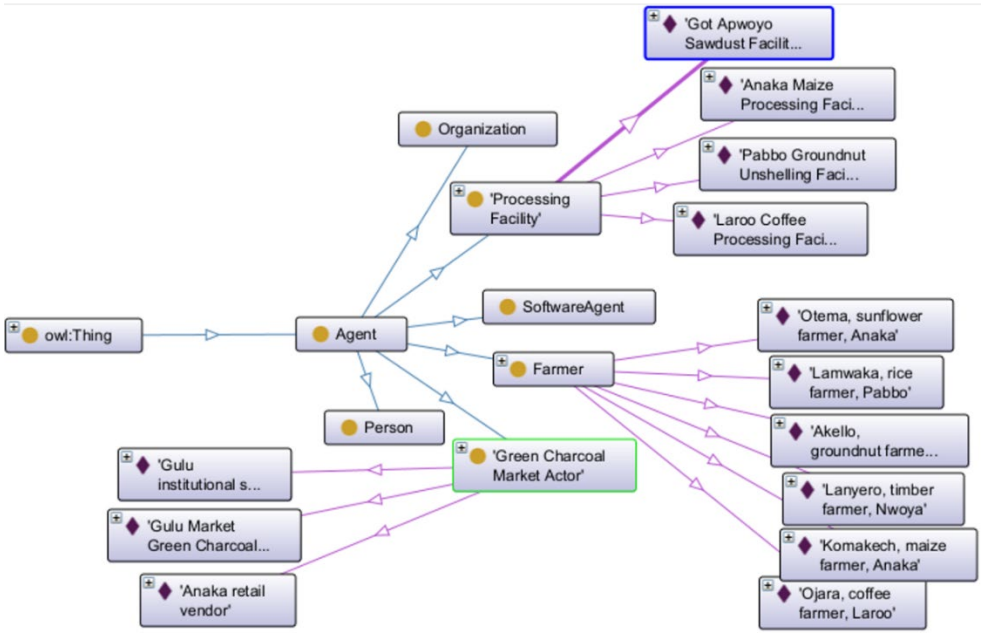


Figure 2c. WP3 Actor Model: Farmer, ProcessingFacility, and GCMarketActor as separate prov: Agent subclass

WP3 (Market and Livelihood) (Figure 2c) was modeled using a role-based agent structure. Farmer, ProcessingFacility, and GCMarketActor were defined as subclasses of prov: Agent to distinguish production, processing, and market roles. ProcessingTransaction was represented as a prov: Activity, while SalesRecord and SustainabilityRecord were modeled as *prov: Entity* instances to capture transaction and sustainability information. MarketSector was implemented as a controlled vocabulary with predefined individuals (Household, School, Restaurant, and Institution) to support consistent querying and reasoning. As the semantic bridge between WP1 and WP5, WP3 implements cross-module properties including suppliesReceipt, resultsInReceipt, locatedInZone, and tracksSalesOf. The module was populated with representative instances of farmers, processing facilities, and market actors for reasoning, validation, and competency-question evaluation. WP5 (Geospatial module) (Figure 2b) is centered on BiomassZone, modeled as a subclass of GeoSPARQL's geo: Feature to support standardized spatial semantics and querying. Environmental observations, represented by *ClimateRecord, AgricultureProductionRecord,* and *EnergyAssessment*, are modeled as subclasses of prov: Entity and linked to biomass zones via the semantic object properties *hasClimateRecord* and *hasProductionRecord*. Cross-module interoperability is achieved through the correspondsToReceipt property, which connects energy assessment records with production data in WP1. Although ENVO was evaluated for environmental representation, it introduced unnecessary semantic complexity and significantly increased reasoning overhead. A lightweight environmental ontology tailored to the study requirements was therefore developed, providing the required functionality while maintaining computational efficiency. The module was populated with four BiomassZone instances and six EnergyAssessment instances used during ontology evaluation.

8. Evaluation

The GCO was evaluated through a DSRM artifact evaluation process, as an experimental validation approach using the HermiT OWL reasoner within Protégé 5.6.9. The complete ontology and its imported modules were successfully loaded in 7.32 s, while ontology classification and inference computation required 1.27 s. The reasoning process pre-computed six categories of inferences, including class hierarchy, object property hierarchy, data property hierarchy, class assertions, object property assertions, and same-individual reasoning, without reporting logical inconsistencies. These results demonstrate that the proposed ontology can be efficiently classified within the proof-of-concept implementation, thereby supporting semantic consistency and interoperability. Given the multidimensional nature of ontology quality, a multi-method evaluation strategy was adopted to assess four key dimensions: logical correctness, data quality, constraint compliance, and functional knowledge retrieval with semantic interoperability. Logical consistency and inference capabilities were evaluated using OWL 2 DL reasoners, specifically HermiT, which supports expressive OWL constructs crucial to the ontology's implementation. Other reasoning engines, such as Pellet, FaCT++, and ELK, were explored, but HermiT was chosen for tasks such as classification and detection of unsatisfiable classes. Data quality was verified using SHACL with pySHACL, in which domain constraints were encoded as SHACL shapes, and RDF instances were assessed against these shapes to identify non-conforming records, particularly those that violated predefined quality requirements. The functional adequacy and knowledge retrieval capabilities were assessed through DL Query within Protégé and SPARQL queries based on competency questions. DL Query was used to verify ontology-level retrieval and inferred knowledge. At the same time, competency questions derived from domain requirements were translated into SPARQL queries to evaluate whether integrated knowledge across WP1, WP3, and WP5 could be retrieved through the defined semantic relationships. These experimental evaluation methods provided evidence that the GCO artifact satisfied its intended objectives by maintaining logical consistency, enforcing domain constraints, and enabling semantic integration and retrieval across interdisciplinary research domains.

9. Communication

This phase disseminates the developed GCO through this publication to researchers, ontology engineers, and practitioners involved in interdisciplinary research data management. The paper presents the ontology specification, design, implementation, and evaluation, demonstrating its ability to integrate heterogeneous Green Charcoal research data through a shared semantic model. By documenting the ontology engineering process and validation results, this publication contributes a reusable ontology artifact that supports semantic interoperability and provides a foundation for similar interdisciplinary research domains.

IV. RESULTS

1. Ontology Structural Integrity

Before evaluating logical consistency, the integrated ontology's structural metrics were meticulously inspected using Protégé 5.6. This step was crucial to confirm that the three work-package modules merged faultlessly without namespace conflicts or declaration errors. The merged ontology comprises 2,674 axioms, including 979 logical axioms, 73 classes, 91 object properties, 103 data properties, and 89 individuals. The absence of conflicting declarations and unintended equivalence relationships indicates that the imported modules were successfully integrated into a unified ontology network. This structural integrity is essential to ensuring the ontology functions effectively across various research domains, as it maintains semantic coherence and facilitates interoperability among diverse datasets. To further validate the ontology's logical consistency, it was processed using the HermiT reasoner, which completed the evaluation in 1373 milliseconds (ms). This rapid processing time demonstrates the efficiency of the ontology's design and the effectiveness of HermiT in handling

complex reasoning tasks. The successful classification of the merged ontology without structural conflicts or unsatisfiable classes reinforces the reliability of the GCO as a robust semantic model for interdisciplinary research.

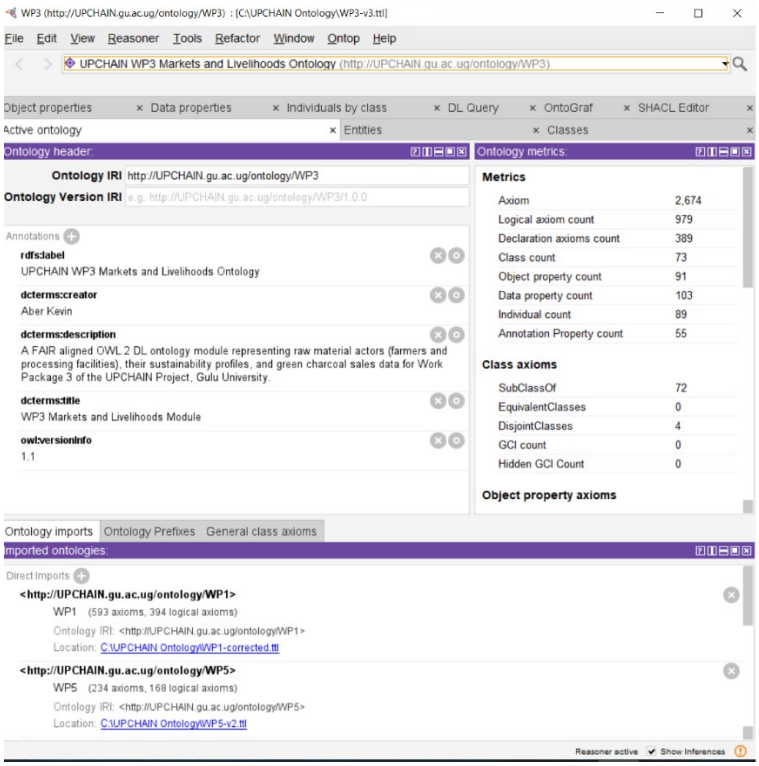


Figure 3. Ontology metrics of the merged WP3 ontology module importing WP1 and WP5.

2. SHACL Constraint Validation

While OWL reasoning effectively verifies the ontology's semantic consistency, it does not enforce instance-level quality constraints due to the open-world assumption inherent in OWL. To address this limitation, SHACL validation was employed using pySHACL to assess the compliance of populated RDF instances against specific domain constraints. In this evaluation, six CarbonizationProcess individuals were scrutinized against the defined SHACL shapes. The validation results indicated a "Conforms: False" status, revealing two instances: CarbProc_ANK_006_B01 (16.9%) and CarbProc_PAB_003_B01 (18.5%) that exceeded the established moisture-content threshold of 15%, as in Figure 4a. This finding is significant because it highlights potential quality issues in the data that could affect downstream applications and decision-making in the green charcoal production sector. To further validate the consistency of the encoded quality rules, the moisture constraint was independently assessed using a DL Query executing in 132 ms, as shown in Figure 4b. The query expression: *{CarbonizationProcess} and (moistureContentAfter some } xsd:decimal[<=15.0])* returned four of the six CarbonizationProcess individuals: *CarbProc_ANK_002_B01, CarbProc_GUL_001_B01, CarbProc_GUL_004_B01, and CarbProc_NWO_005_B01*. Notably, these results correspond precisely to the complement of the two instances identified by SHACL validation, reinforcing the reliability of the quality constraints. The agreement between the findings from OWL-based reasoning and SHACL validation provides convergent evidence that the quality constraints were accurately represented at both the semantic and data validation levels. This dual validation approach not only enhances the ontology's robustness but also ensures it

effectively captures and enforces critical quality parameters necessary to maintain data integrity in interdisciplinary research contexts. The identification of non-conforming instances serves as a crucial feedback mechanism for refining the ontology and improving data quality management practices.

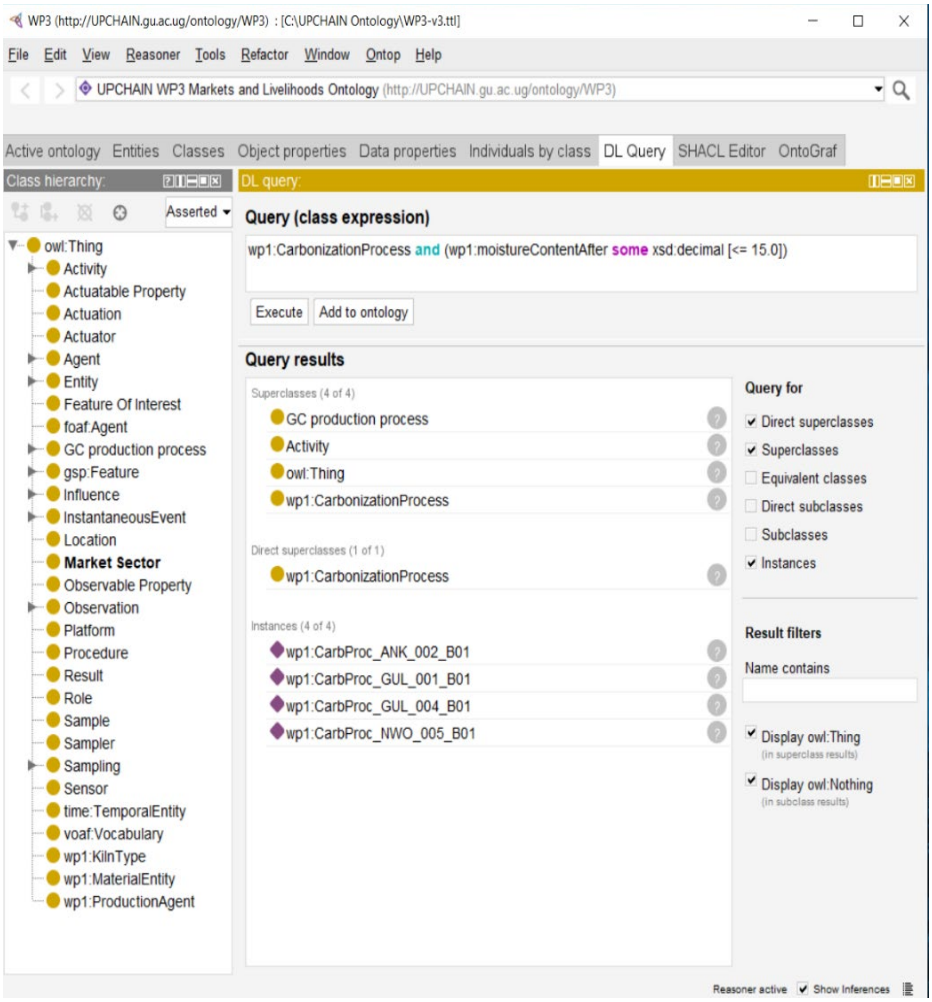


Figure 4a. pySHACL Validation report for the populated WP1 Module

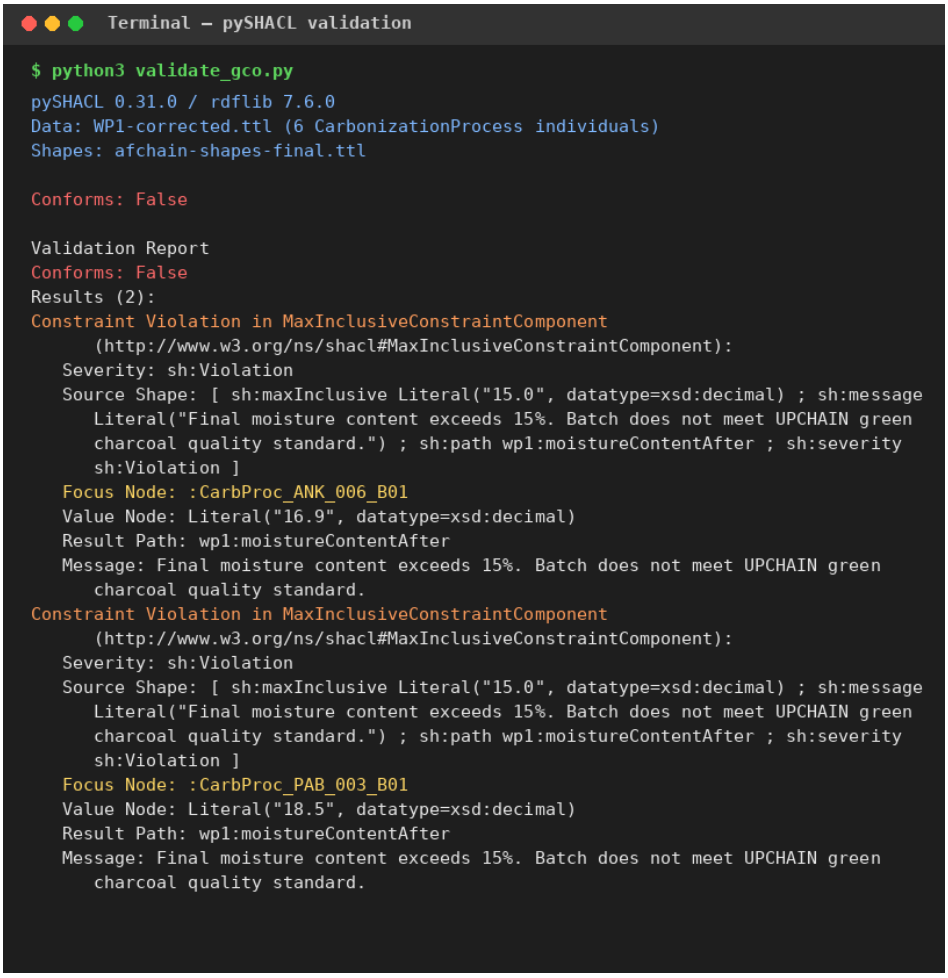


Figure 4b. DL Query Cross-Check in Protege, confirming the quality conformant batches identified by SHACL

3. Competency Question and Cross-Domain Retrieval Evaluation

The functional capability of the Green Charcoal Ontology (GCO) was rigorously assessed through competency questions derived from a comprehensive analysis of domain requirements. These questions were translated into SPARQL queries and executed against a populated knowledge graph containing 1,343 triples from Work Packages 1, 3, and 5, as illustrated in Figure 5 (a & b). CQ1 aimed to determine whether the ontology could effectively trace the complete chain of a batch from feedstock to market. The results, depicted in Figure 5a, revealed six unique entries, each detailing specific associations among feedstock, production agents, moisture readings, and market sectors. The absence of repetitions in the provenance chains indicates that the ontology successfully preserves per-batch provenance. This finding confirms that the GCO accurately links production entities to market actors while preserving distinct details, thereby meeting the structural coverage criteria essential for effective data management.

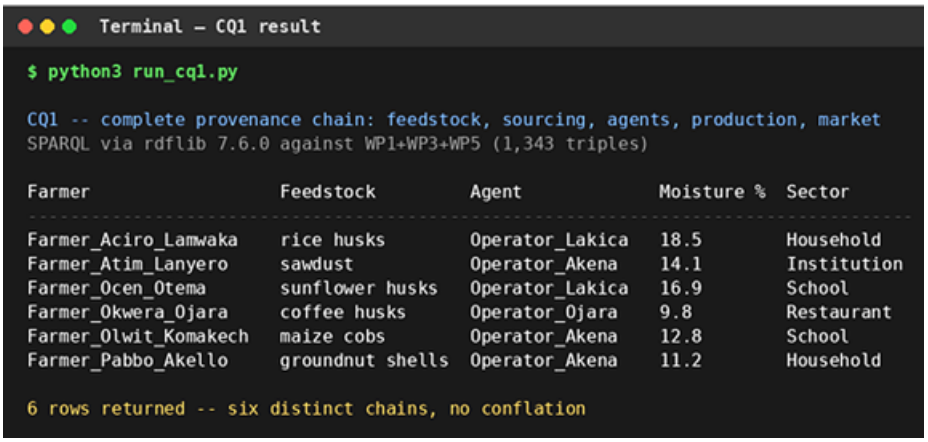


Figure 3. CQ1 Result: Six Distinct Provenance chains, one per farmer, each with a different feedstock, agent, and market sector

CQ2 focused on evaluating the economic relationship between batch quality and market outcomes as represented in the ontology. The results highlighted in Figure 5b demonstrated a clear trend: as moisture content increased, the calorific value decreased, leading to declines in both selling price and quantity sold. Notably, the two batches labeled as "Rejected," with moisture contents of 16.9% and 18.5%, recorded the lowest selling prices and quantities. These batches were also identified as SHACL violations, which were excluded from the DL query results. This consistency reinforces the ontology's capability to identify out-of-specification batches across multiple evaluation layers, ensuring that quality control measures are effectively integrated into the data management process.



Figure 4. CQ2 Result: Selling Price and Quantity sold fall as quality falls, rejected batches sit at the lowest

Throughout the evaluation, the two batches, CarbProc_ANK_006_B01 and CarbProc_PAB_003_B01, were consistently flagged as non-conforming quality instances through various validation methods, including SHACL constraint validation, OWL 2 DL class-expression reasoning, and SPARQL-based competency queries. Furthermore, the merged ontology was successfully classified, with no structural conflicts or unsatisfiable classes detected by the reasoner. The convergence of results from these complementary evaluation approaches provides robust evidence of the ontology's validity. Each method examines distinct quality dimensions: structural

coherence, logical entailment, data-constraint compliance, and functional knowledge retrieval, demonstrating that the GCO is a logically consistent and interoperable semantic model for interdisciplinary green charcoal research. Through this modular ontology design and reusing established Semantic Web vocabularies, the GCO operationalizes FAIR principles, significantly enhancing the findability, accessibility, interoperability, and reuse of heterogeneous research knowledge. This comprehensive evaluation underscores the ontology's potential to facilitate effective data integration and management in complex research environments, ultimately supporting informed decision-making in the green charcoal sector.

V. DISCUSSION

This study highlights the significance of ontology-driven semantic mediation in enhancing interdisciplinary research integration, particularly through the Green Charcoal Ontology (GCO). The findings underscore that effective semantic interoperability transcends technical data exchange, focusing instead on coordinating meaning across diverse knowledge practices. The GCO's successful integration of production, market, and geospatial concepts illustrates that achieving interoperability does not require replacing existing data infrastructures or enforcing complete conceptual agreement across disciplines. Instead, it establishes a shared semantic layer that connects independently managed datasets while preserving the autonomy of domain-specific practices. The GCO functions as a formal semantic boundary object, facilitating collaboration among research communities while allowing each domain to retain its unique interpretations and practices. This perspective aligns with Star and Griesemer's [20] concept of boundary objects, emphasizing that ontology development is not merely a computational modeling task but a process of negotiating shared meaning across different knowledge systems. This is particularly crucial in sustainability research, where integrating scientific measurements, spatial observations, stakeholder experiences, and socio-economic knowledge is essential to addressing complex research questions. A key implication of this study concerns the representation of knowledge from qualitative and contextual research. Unlike many ontology-based integration approaches developed around structured quantitative datasets with relatively stable schemas, UPCHAIN involves knowledge generated through field practices, stakeholder interactions, and domain-specific interpretations. Such knowledge cannot always be captured through direct schema transformation. The development of GCO therefore required practice-informed concept identification, competency questions, vocabulary alignment, and iterative modeling to capture contextual meanings while maintaining machine interpretability [40–42]. This demonstrates the importance of grounding ontology engineering in the practices through which knowledge is produced. This practice-informed approach supports Carlile's view that interdisciplinary collaboration depends on exchanging and transforming knowledge, rather than merely transferring data [43]. The study further contributes to discussions on global versus modular ontology architectures. While universal ontologies aim to achieve broad conceptual consensus, complete alignment across interdisciplinary research communities is difficult because disciplinary assumptions, research objectives, and local contexts shape concepts. Conversely, highly specialized ontologies may provide detailed representation but limited interoperability beyond their original domain. The modular hybrid architecture adopted in GCO addresses this tension by maintaining work-package-specific concepts while establishing semantic bridges through shared concepts and reused Semantic Web vocabularies. This demonstrates that interoperability does not require identical conceptual models but mechanisms that enable meaningful alignment between heterogeneous perspectives [25]. Beyond semantic integration, the GCO serves as the semantic blueprint for the Research Data Management System (RDMS). It defines core entities, relationships, provenance structures, and validation constraints that guide system implementation while remaining independent of specific technologies. Unlike conventional integration approaches that primarily rely on structural mappings or relational joins, the ontology explicitly represents domain meaning and enables reasoning, provenance tracking, and future system extensions. The combined use of OWL reasoning, SHACL validation, and competency-question evaluation provides a layered quality assurance approach that assesses logical consistency, data conformity, and practical usability [44]. This reinforces the view that FAIR implementation requires more than metadata availability; it requires machine-interpretable representations of concepts, relationships, and provenance [10], [11], particularly in research environments where heterogeneous

systems must continue to operate independently [45]. The practical impact of the ontology extends to multiple stakeholder groups, including researchers, data managers, GIS specialists, field researchers, and future data users. Before GCO, datasets generated by different work packages required substantial manual interpretation and reconciliation before cross-domain analysis could be performed. The ontology enables researchers to query relationships between biomass resources, production characteristics, spatial distributions, and livelihood outcomes within a shared semantic environment. Thus, its contribution is not only the formal representation of domain concepts but also the creation of an infrastructure that supports interdisciplinary knowledge discovery and data reuse. By reusing established Semantic Web vocabularies, including SOSA, PROV-O, and GeoSPARQL, locally developed research concepts align with international standards while preserving contextual and cultural knowledge relevant to the project. This makes the framework transferable to similar interdisciplinary and resource-constrained research environments. This approach is particularly relevant for resource-constrained research environments where replacing existing infrastructures may be impractical. By introducing a semantic layer over existing systems, ontology-driven mediation provides an incremental pathway toward FAIR-aligned research ecosystems. This is valuable in African research contexts, where fragmented infrastructure, limited technical capacity, and isolated data repositories remain barriers to effective reuse of research data [35], [36]. The study demonstrates that semantic interoperability can be progressively achieved without disrupting established data collection practices. Several limitations remain. The evaluation was conducted as part of a single interdisciplinary research project, and wider adoption across institutions and domains requires further validation. In addition, ontology reuse required balancing semantic richness, reasoning performance, and maintainability. Future work will investigate broader stakeholder validation, larger-scale datasets, automated FAIR assessment, and integration within federated research infrastructures. Nevertheless, the findings demonstrate that practice-informed, modular ontology engineering provides a practical and scalable approach for enabling FAIR-aligned semantic interoperability while preserving the contextual meanings embedded within interdisciplinary research practices.

VI. CONCLUSION

This study presented the Green Charcoal Ontology (GCO), a FAIR-aligned, ontology-driven semantic mediation framework to improve interoperability across heterogeneous interdisciplinary research data environments. Using the UPCHAIN project as an empirical case, the framework integrates production, market, and geospatial knowledge through a practice-informed ontology that aligns local research concepts with established Semantic Web standards, including SOSA, PROV-O, and GeoSPARQL. The evaluation demonstrated that GCO supports logical consistency, data quality validation, and cross-domain knowledge discovery through OWL reasoning, SHACL validation, and competency-question-based SPARQL queries. Beyond enabling semantic integration, the GCO provides a conceptual blueprint for the Research Data Management System (RDMS) by defining shared concepts, relationships, provenance structures, and validation mechanisms required for interoperable research workflows. The findings highlight that achieving FAIR interoperability requires more than standardized metadata; it requires explicit representation of meaning, context, and relationships across research domains, positioning ontology development as both a technical and practice-grounded process. This work demonstrates how semantic models can support collaboration among diverse research communities while preserving domain-specific knowledge practices. Future work will investigate broader stakeholder validation, deployment across additional research domains, automated FAIR assessment, and integration within federated research data infrastructures. The proposed approach provides a practical foundation for advancing semantic interoperability and the development of reusable research data ecosystems, particularly in interdisciplinary and resource-constrained environments.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the UPCHAIN project team for their valuable contributions throughout this research. We also thank the youth group in Pabbo for their participation in the field experiments and data collection that informed the development of the ontology. This work was financially supported by the Ministry of Foreign Affairs of Denmark.

REFERENCES

[1] Chamanara J, Kraft A, Auer S, et al. Towards Semantic Integration of Federated Research Data. *Datenbank-Spektrum* 2019; 19: 87–94.
[2] Campos VV de S, Brancher JD, Farias FP, et al. Review and comparison of works on heterogeneous data and semantic analysis in Big Data. *Semin Ciências Exatas e Tecnológicas* 2021; 42: 113.
[3] Bakshi WJ. Leveraging Semantic Technologies in ETL Processes for Data Integration in Heterogeneous Environments. *Educ Adm Theory Pract* 2024; 27: 1324–1328.
[4] Calvanese D, Lanti D, Mendes De Farias T, et al. Accessing scientific data through knowledge graphs with Ontop. *Patterns* 2021; 2: 100346.
[5] Carlos MJ, Baptista AA, Sousa RM, et al. Semantic Mediation: a literature review on semantic interoperability through ontologies. *Procedia Comput Sci* 2025; 263: 734–743.
[6] Campos LM. Classificação de objetos de fronteira na organização do conhecimento e o papel das ontologias. *Liinc em Rev* 2018; 14: e4314.
[7] Huvila I. Boundary Objects in Information Science. *J Assoc Inf Sci Technol* 2017; 68 (8): 1807–1822.
[8] Fusco G, Aversano L. An approach for semantic integration of heterogeneous data sources. *PeerJ Comput Sci* 2020; 2020: 1–30.
[9] Calvanese D, Gal A, Lanti D, et al. Conceptually-grounded mapping patterns for Virtual Knowledge Graphs. *Data Knowl Eng* 2023; 145: 102157.
[10] Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016; 3: 1–9.
[11] Annika Jacobsen1, Ricardo de Miranda Azevedo2 NJ, Dominique Batista4 SC, Cornet6 R, et al. FAIR Principles: Interpretations and Implementation Considerations Annika. *Data Intell* 2019; 23: 0–2.
[12] David R, Baumann K, Franc Y Le, et al. Converging on a semantic interoperability framework for the European Data Space for Science, Research and Innovation (EOSC), https://documentserver.uhasselt.be/handle/1942/41709%0Ahttps://documentserver.uhasselt.be/bitstream/1942/41709/1/SITeaserPaper_FOIS2023.pdf (2023).
[13] Kamdar MR, Musen MA. An empirical meta-analysis of the life sciences linked open data on the web. *Sci Data* 2021; 8: 1–21.
[14] Oliveira D, Pesquita C. Improving the interoperability of biomedical ontologies with compound alignments. *J Biomed Semantics* 2018; 9: 1–13.
[15] Peng C, Xia F, Naseriparsa M, et al. Knowledge Graphs: Opportunities and Challenges. *Artif Intell Rev* 2023; 56: 13071–13102.
[16] Torab-Miandoab A, Samad-Soltani T, Jodati A, et al. Interoperability of heterogeneous health information systems: a systematic literature review. *BMC Med Inform Decis Mak* 2023; 23: 1–13.
[17] Palojoki S, Lehtonen L, Vuokko R. Semantic Interoperability of Electronic Health Records: Systematic Review of Alternative Approaches for Enhancing Patient Information Availability. *JMIR Med Informatics* 2024; 12: 1–11.
[18] Carlile PR. A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organ Sci* 2002; 13: 442–455.
[19] Li J, Xian G, Zhao R, et al. RDFAdaptor: Efficient ETL Plugins for RDF Data Process. *J Data Inf Sci* 2021; 6: 123–145.
[20] Star SL, Griesemer JR. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and

Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Soc Stud Sci* 1989; 19: 387–420.
[21] Corcho O, Priyatna F, Chaves-Fraga D. Towards a new generation of ontology-based data access. *Semant Web* 2020; 11: 153–160.
[22] Gu Z, Lanti D, Corcoglioniti F, et al. Ontology-based data federation and query optimization. *Knowledge-Based Syst* 2025; 329: 114216.
[23] Arenas-Guerrero J, Pérez MS, Corcho O. LUBM4OBDA: Benchmarking OBDA Systems with Inference and Meta Knowledge. *J Web Eng* 2023; 22: 1163–1186.
[24] Buron M, Goasdoué F, Manolescu I, et al. Ontology-based RDF integration of heterogeneous data. *Adv Database Technol - EDBT* 2020; 2020-March: 299–310.
[25] Xiao G, Ding L, Cogrel B, et al. Virtual knowledge graphs: An overview of systems and use cases. *Data Intell* 2019; 1: 201–223.
[26] Fernandez M, Perez AG, Juristo N. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. *Proc - 3rd Int Conf Autom Prod Cross Media Content Multi-Channel Distrib AXMEDIS 2007* 1997; 115–122.
[27] Abu-salih B. MetaOntology: Toward developing an Ontology for the Metaverse. 2022; 1–10.
[28] Suárez-Figueroa MC, Gómez-Pérez A, Fernández M. The N e O n Methodology framework : A scenario-based methodology for ontology development.
[29] Barba-González C, García-Nieto J, Roldán-García M del M, et al. BIGOWL: Knowledge-centered Big Data analytics. *Expert Syst Appl* 2019; 115: 543–556.
[30] Kreuzthaler M, Brochhausen M, Zayas C, et al. Linguistic and ontological challenges of multiple domains contributing to transformed health ecosystems. *Front Med* 2023; 10: 1–15.
[31] Campos PMC, Reginato CC, Almeida JPA, et al. Finding reusable structured resources for the integration of environmental research data. *Environ Model Softw*; 133. Epub ahead of print 2020. DOI: 10.1016/j.envsoft.2020.104813.
[32] Weber M, Buche P, Ibanescu L, et al. PO2/TransformON, an ontology for data integration on food, feed, bioproducts, and biowaste engineering. *npj Sci Food* 2023; 7: 1–14.
[33] Gaignard A, Rosnet T, De Lamotte F, et al. FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. *J Biomed Semantics* 2023; 14: 1–15.
[34] Islam S. FAIR digital objects, persistent identifiers, and machine actionability. *FAIR Connect* 2023; 1: 29–34.
[35] Chiware ERT. Open research data in African academic and research libraries: a literature analysis. *Libr Manag* 2020; 41: 383–399.
[36] Mosha NF, Ngulube P. The utilization of open research data repositories for storing and sharing research data in higher learning institutions in Tanzania. *Libr Manag* 2023; 44: 566–580.
[37] De Santis A, Balduini M, De Santis F, et al. Integrating Large Language Models and Knowledge Graphs for Extraction and Validation of Textual Test Data. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2025; 15233 LNCS: 304–323.
[38] Iglesias-Molina A, Van Assche D, Arenas-Guerrero J, et al. The RML Ontology: A Community-Driven Modular Redesign After a Decade of Experience in Mapping Heterogeneous Data to RDF. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 2023; 14266 LNCS: 152–175.
[39] Sakor A, Brunet M, Iglesias E, et al. Integrating Knowledge Graphs and Neuro-Symbolic AI: LDM Enables FAIR and Federated Research Data Management. *WSDM 2025 - Proc 18th ACM Int Conf Web Search Data Min* 2025; 1044–1047.
[40] Hocker J, Schindler C, Rittberger M. Participatory design for ontologies: a case study of an open science ontology for qualitative coding schemas. *Aslib J Inf Manag* 2020; 72: 671–685.
[41] Keet CM, Khan ZC. Characterizing Competency Questions for Ontologies. *CEUR Workshop Proc*; 4176.



[42] Raza MA, Rahmah M, Raza S, et al. A methodology for Engineering domain ontology using entity relationship model. *Int J Adv Comput Sci Appl* 2019; 10: 326–332.
[43] Carlile PR. A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organ Sci*; 13. Epub ahead of print 2002. DOI: 10.1287/orsc. 13.4.442.2953.
[44] Oudshoorn A, Ortiz M, Šimkus M. SHACL validation in the presence of ontologies: Semantics and rewriting techniques. *Artif Intell* 2026; 352: 104483.
[45] Welter D, Juty N, Rocca-Serra P, et al. FAIR in action - a flexible framework to guide FAIRification. *Sci Data* 2023; 10: 1–9.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Comparative Analysis of Optimal Curtailment Strategies for Distributed Energy Resources

Mustafa Aker
Research and Development
Siemens A.S.
Istanbul, Turkey
mustafa.aker@siemens.com
ORCID: 0000-0001-8972-1020

Abstract
The rapid growth of distributed energy resources in modern grid topologies brings significant operational challenges, including transformer overloads and voltage constraint violations caused by unexpected generation and consumption across the grid. Resolving these incidents requires curtailment allocation algorithms that coordinate the available flexibility of DERs according to prioritized criteria. This paper presents a formal comparative analysis of four curtailment allocation algorithms from an active network management perspective. Equal-percentage curtailment enhanced by a water-filling mechanism, capacity-ranked minimum DER-count greedy curtailment, merit-order curtailment, and grid-aware curtailment methodologies have been compared within this DER management framework. A benchmark scenario involving nine PV-based DERs was constructed using IEEE 123-bus QSTS feeder data from the Open Energy Data Initiative to reveal differences among these four strategies. According to the results, no single strategy dominates across all other methodologies. The most suitable choice for different situations depends on whether the system operator prioritizes fairness, operational simplicity, or economic efficiency, as well as on contract types. Rather than proposing a universally optimal allocation method, this study provides a transparent framework for selecting a curtailment strategy based on the operator's priorities and available grid information.

Keywords: greedy algorithms, distributed energy resources, DER curtailment, active network management, network constraint optimization, smart grid

I. INTRODUCTION

The global energy transition in grid networks has progressed rapidly in recent years. This growth mainly stems from the increasing number of distributed energy resources across all grid layers, including low-, medium-, and high-voltage layers. Electric vehicle chargers, roof photovoltaic systems, and heat pumps are becoming more widespread day by day. With this change, the conventional grid topology is evolving, shifting the unidirectional power-flow characteristic from generator-side equipment to consumer-side equipment [1]. As DER penetration in grid networks increases, operators at grid operations centers face increasingly challenging situations in maintaining grid stability while allocating these DERs within their operational boundaries. Reverse power flow, transformer overload, and congestion among feeders are the most frequent effects of the increasing number of DERs [2]. Conventional grid management methodologies, such as open-close switches, are inadequate for this kind of new-generation grid network enriched with distributed energy resources. At this very point, the active network management paradigm stepped onto the stage with the aim of solving these types of problems. Instead of changing the network topology as seen in traditional approaches, this approach benefits from real-time monitoring and resource control over grid systems, while managing these distributed energy systems within their operational limits based on congestion [1]. In this manner, grid assets are continuously monitored for their measurement values, and control operations are determined in control rooms using software systems. These

operations are decided, calculated, and sent to grid systems to relieve the congestion in grid networks [3]. Recent research has offered a congestion management perspective that includes direct DER control, local flexibility markets, and flexible tariffs, all of which introduce different operational requirements and trade-offs [4],[5]. The power of active network management rises from the ability to modulate DER output. This modulation applies to both generation and consumption units. This requires both awareness of device capacities and evaluation of each device's topological position within the grid network and its influence at the grid level [6],[7],[8]. When a grid network violation is detected, the operator needs to determine which DERs to curtail and how much to curtail them. This curtailment allocation problem can be addressed using various algorithmic approaches that address different criteria, such as operational objectives, optimality criteria, and correctness factors. The present paper focuses on the design and evaluation of curtailment allocation strategies for heterogeneous DER pools. Specifically, four different greedy algorithms, which differ in their optimization objectives, data requirements, and behavioral properties, are used for the curtailment allocation problem. Compared with network-constrained power flow methodologies, which require detailed network models and may impose significant computational and data management requirements, the evaluated approaches rely on sorting and simple iterative calculations [5],[9]. Thanks to sorting-based implementations, analyzed algorithms can operate with O n log n), O (n), or O (n) sağ parantez time complexity, where n represents the number of DERs connected to the grid. Grid operators may choose different strategies for congestion cases, such as prioritizing fairness across DERs, minimizing the number of affected equipment, optimizing operational cost efficiency, and applying the operation that provides the greatest relief at a specific location on the grid. These strategies have different consequences for DER owner experience, system economics, and grid operations. Recent fairness-focused studies show that fairness, economic benefits, and network effectiveness may lead to different allocation outcomes and may not be optimized simultaneously [10],[11]. The analysis presented in this work aims to provide a structurally and practically grounded comparison among these strategies to inform decisions on the design and configuration of smart grid management platforms. Related to this, the main contribution of this work is not the introduction of a new superior algorithm, but the comparison of four perspectives: proportional fairness, minimum DER activation, economic priority, and electrical effectiveness.

II. RELATED WORK

Over the past two decades, energy grids have evolved to include more distributed energy resources, and the challenges of this process have begun to rise. To prevent congestion in grid networks, curtailment of DERs is treated as the last resort by early active network management systems; only generation-side curtailments were applied as a solution, since they are seen as more acceptable from a cost perspective [1],[12]. Later adoptions were directed toward the consumption-side flexibility and voltage regulation operations, which provide more comprehensive efficiency and fairness in real-time network management [13]. More recent research classifies distribution-level congestion management into direct control mechanisms, local flexibility markets, and tariff-based approaches. The applicability of these mechanisms is highly dependent on network conditions, local regulations, available information, and client-side objectives [4]. According to the current active network management literature, curtailment allocation strategies can be evaluated from two different perspectives: optimization and computational differences. Proportional curtailment, in which each DER connected to the grid contributes the share of the curtailment amount based on its proportional availability, due to its simplicity and fairness [14]. However, this solution should be supported by a water-filling mechanism to ensure feasibility by iteratively redistributing the remaining curtailment amount among the DERs that still have availability to curtail. Recent fairness-oriented research has shown that equal, proportional, and dynamic allocation approaches can yield different curtailment distributions even when they satisfy the same network constraints [10]. Fairness-based methodologies can prevent situations in which the system repeatedly curtails a specific DER owner. However, regarding fairness-based strategies, this approach may involve trade-offs that result in higher total

system curtailment while compromising economic efficiency [11]. Another possible operational approach is to minimize the number of DERs involved during curtailment operation. In this study, this objective is represented by a capacity-ranked minimum DER count greedy methodology that curtails higher-capacity DERs first. This approach can reduce the number of control commands affecting grid assets and, accordingly, contractual interactions. However, it does not take into account the fairness and cost-effectiveness of each DER, so it is treated as an operational-simplicity heuristic rather than an optimal congestion-management strategy. The merit order curtailment approach adopts the perspective of curtailing the cheapest resources first, once available, and provides a cost-optimal solution for congestion cases [15]. When partial curtailment is allowed according to clients' contract types, the allocation problem begins to resemble a fractional resource allocation problem rather than the classical knapsack problem [16]. Ordering DERs based solely on their curtailment costs is cost-optimal only when all curtailed units provide the same amount of violation relief. In cases where electrical sensitivities differ, a cost-to-relief metric may be required. In that sense, more complex market-based approaches incorporate congestion pricing, demand-response preferences, and client-based constraints into their methodologies [17]. While these strategies provide a comprehensive representation of economic behavior, they should be informed by additional market, contractual, and operational data, unlike a simple merit order mechanism. Grid-aware curtailment is a solution in which curtailment decisions are made based on the weighted electrical influence of each DER on the grid network. This approach is considered in the context of optimal power flow decomposition and distributed grid control operations within the grid network [18], [19]. To benefit from this approach in real-time solutions, power transfer distribution factors and voltage sensitivity coefficient values need to be prepared offline for efficient calculations, provided network topologies remain the same over long-term executions [6]. Recent voltage sensitivity-based control approaches use active and reactive power sensitivity matrices to identify location-aware effects of DERs and to inform voltage management decisions [7]. Sensitivity coefficients have also been applied to flexibility management, involving both node-voltage and line-loading constraints [8]. In recent studies, comprehensive alternating-current optimal power flow-based distribution systems have begun to emerge, and their flexibility platforms are designed to evaluate network constraints, flexibility offers, and operational schedules simultaneously [5]. These formulations may yield more feasible and optimized decisions based on network and economic criteria, but they heavily rely on detailed grid models and extensive input data on grid assets. In situations lacking such data, model-free reinforcement learning methods have also been proposed to learn fair DER curtailment strategies [20]. Such approaches reduce overall dependence on additional data but require training iterations to provide meaningful operational predictability. Existing studies provide advanced solutions for individual objectives such as fairness, market efficiency, network feasibility, and model-free perspectives; however, this study simultaneously compares four operational principles: proportional fairness, minimum DER activation, merit order cost priority, and sensitivity-based grid effectiveness. It is described how grid operators' preferred allocation strategy changes depending on the objective and the available data on grid networks.

III. EXPERIMENTAL STUDY & ANALYSIS

1. System Model

Within the scope of this study, the distribution network monitored by an active network management framework is considered. With that, the operator can monitor real-time telemetry data from a grid network with n flexible distributed energy resources connected. Each DER also has set of operational parameters such as its current output a_i minimum opratioal limits min_i and maximm ratd apacity max_i. Full notation use throughout the paper is given in Table 1. The allocation problem is evaluated for a single operational perspective, where the curtailment decision is made. The corresponding active network management framework performs the identification of single eligible DERs and calculation of the required relief. This study focuses specifically on

the subsequent allocation of requirements among the eligible DERs. For grid-aware curtailment, sensitivity coefficients are assumed to remain valid during the allocation interval and to be recalculated upon physical changes in network topology or operating conditions.

2. Violation Detection and Curtailment Trigger

Thanks to the active network management framework, once a constraint violation is detected for a specific network element, the grid operator can trigger a curtailment event. These violations may be line or transformer overload that corresponds to the network.

Table 1. Variables and variable descriptions used in the current study

Variable Name	**Description**
n	Number of flexible DER assets
a_i	Current analog value of DER i
min_i	Minimum allowed operating value of DER i
max_i	Maximum rated value of DER i
k_i	Curtailability factor of DER i, $k_i \in (0,1]$
h_i	Available curtailment amount, $h_i = a_i - min_i$
q_i	Maximum realizable curtailment, $q_i = k_i \cdot a_i - min_i$
c_i	Curtailment cost of DER i
s_i	Electrical sensitivity coefficient of DER i with respect to the violated element, $s_i \in (0,1]$
C_{req}	Total curtailment amount
x_i	Curtailment amount assigned to DER i, $0 \leq x_i \leq q_i$
I_{req}	Required weighted electrical relief for the violated network element in the grid-aware formulation

element exceeds its rated capacity, or overvoltage events in network elements that exceed the upper bounds defined by grid constraints. In such cases, the grid operator computes the required aggregate curtailment amount, Calt req, and then provides it as input to curtailment allocation algorithms. Calculation of the active network management framework and kept identical for all aggregate-based curtailment algorithms. This separation allows us to compare only allocation decisions produced by the algorithms under the same detected violation. For the grid-aware curtailment algorithm, the operator provides the required impact value I alt req, which is derived from the magnitude of the detected violation. Resunokta s alt i, büyük veya eşittir I alt req , son toplama, where the left-hand side represents cumulative weighted electrical relief provided for violated network equipment.

3. Curtailment Allocation Problem

Upon detection of a violation incident and the required curtailment amount, the Calt req is calculated by the operator. The curtailment allocation problem is addressed by determining a vector of curtailment assignments x eşittir (x alt 1,x alt 2 ,…, x alt n)which satisfies the following feasibility constraints.

$$\sum_{i=1}^{n} x_i \geq C_{req}, \quad 0 \leq x_i \leq q_i, \quad \forall i \in \{1,\ldots,n\} \tag{1}$$

First constraint within the (1) ensures that total curtailment is sufficient, so that it can solve the violation, and the second constraint ensures each DER is curtailed inside its physical and contractual boundaries.

In the grid-aware algorithm, aggregate curtailment is replaced by a location-specific feasibility condition.

$$\sum_{i=1}^{n} x_i \cdot s_i \geq I_{req}, \quad 0 \leq x_i \leq q_i, \quad \forall i \in \{1,\ldots,n\} \tag{2}$$

Regarding to (2), aggregate-curtilment instance is feasible when $\sum_{i=1}^{n} q_i \geq C_{req}$ while, grid-aware instance is feasible when $\sum_{i=1}^{n} q_i \cdot s_i \geq I_{req}$. If corresponding condition is not satisfied, that means, available DER flexibility is insufficient to resolve the violation and algorithm reports the allocation problem as infeasible.

Along with the feasibility, different types of secondary objectives may apply depending on priorities of grid operator based on contract types such as providing fair curtailment distribution among DERs, minimizing the number of curtailed DERs by reducing communication overhead to the grid, minimizing the total cost of curtailment applied to the grid, and prioritizing DERs whose electrical locations enable the most effective relief to the violated network element. The following allocation obctives represent these operational priorities: proportion utilization distribution atoplama, and available DERs, minimization of $\sum_{i=1}^{n} c_i \cdot x_i$ and minimization of $\sum_{i=1}^{n} x_i$ subject to location-specific relief toplama, subject (2), respectively.

4. Algorithmic Details

This paper analyzes four different curtailment algorithms: equal-percentage curtailment, capacity-based greedy curtailment that minimizes the number of affected DERs, merit-order (cost-based) curtailment, and grid-aware curtailment.

5. Algorithm Selection

Four strategies were predetermined to represent different grid operator objectives and to provide deterministic allocation rules under DER-level constraints. Equal percentage curtailment represents proportional burden sharing; minimum DER count curtailment aims to reduce the number of controlled assets; merit order curtailment prioritizes economic efficiency; and grid-aware curtailment targets DERs according to their electrical effectiveness while relieving violations on grid networks. Strategies also differ in their data requirements. Equal percentage and minimum DER count strategies use DER curtailment limits while merit order and grid-aware solutions require unit cost and electrical sensitivity data, respectively. Considering these, four curtailment methodologies provide a basis for examining how grid operator priorities and data availability affect curtailment strategy selection.

6. Equal Percentage Curtailment

Algorithm gets set of DERs (n) and required curtailment amount C_{req} as an input and outputs per-DER curtailment alocation x_i such that $\sum_{i=1}^{n} x_i = C_{req}$ whenever aggregate available flexibility is sufficient.

Each DER gets a curtailment share based on its availability, h alt i=a alt i eksi alt simge taban , min , son taban , alt i, considering their operational limits, alt simge taban , m i. n , son taban , alt i, and alt simge taban , m a. x , son taban , alt i. As a result of the water-filling approach, each DER fulfilling its available curtailment share in the current iteration is removed from the next iteration. This process terminates when $C_{req}$is are fully met, or all DERs are saturated. The equal percentage curtailment procedure is expressed in Table 2.

7. Minimum DER Count Capacity-Based Greedy Curtailment

Set of DERs n and required curtailment amount C_{req} is provided as input and per-DER curtailment allocation x_i, using fewest possible DERs expected as output.

All DERs are sorted according to their realizable curtailment values q_i in descending order then, algorithm greedily assigns full curtailment to each DER until C_{req} is fully met. Sorting by q_i values first, ensures number of DERs participated to the solution is minimal. Allocation methodology is expressed in Table 3.

Since the last selected DER can be applied partial curtailment, the minimum number of required DERs corresponds to the smallest integer k where sum of the k number of largest q_i value meets or exceeds C_{req}.

Table 2. Equal percentage curtailment pseudo code

$x_i \leftarrow 0$ for every $i \in D$
Active $\leftarrow \{i \in D : q_i > 0\}$
Remaining $\leftarrow C_{req}$
while Remaining $> \varepsilon$ and Active $\neq \emptyset$ do
  $H \leftarrow \sum (i \in \text{Active}) h_i$
  for every $i \in$ Active do
    $\Delta, \leftarrow \min(q_i - x_i, (h_i/H)\,\text{Remaining})$
    $x_i, \leftarrow x_i, + \Delta,$
  end for
  Remaining $\leftarrow$ Remaining $- \sum (i \in \text{Active}) \Delta,$
  Remove each DER satisfying $x_i \geq q_i - \varepsilon$ from Active
end while
if Remaining $> \varepsilon$, report infeasible
return x

Table 3. Minimum DER count capacity based greedy curtailment pseudo code

For every $i \in D$, calculate $q_i \leftarrow k_i \cdot a_i - \min_i$
Remove every DER satisfying $q_i \leq 0$
Sort the remaining DERs by q_i in descending order

```
Remaining ← C_req
x_i ← 0 for every i ∈ D
for every i ∈ D in sorted order do
   if Remaining ≤ 0, break
   x_i ← min(q_i, Remaining)
   Remaining ← Remaining − x_i
end for
if Remaining > 0, report infeasible
return x
```

8. Merit Order (Cost-Weighted) Curtailment

Algorithm expects set of DERs n, curtailment costs of each DER c_i and required curtailment amount C_{req} value as inputs and gives per DER curtailment allocation x_i, minimizing total curtailment cost $\sum_{i=1}^{n} c_i \cdot x_i$ as an output.

All DERs are sorted according to their curtailment costs c_i in ascending order then, algorithm greedily assign full curtailment to each DER until C_{req} is fully met. Since partial curtailment is allowed with this methodology, this procedure follows continuous fractional allocation rule. Basically, this approach minimizes the direct curtailment cost for the aggregate target, assuming that each curtailed unit provides an equivalent amount of violation relief. Merit order curtailment algorithm is given in Table 4.

Sorting DERs only regarding to their c_i values are not enough for minimizing location specific relief costs under varying electrical sensitivities. Therefore, present merit order algorithm is kept separate from grid-aware formulation as it is evaluated against the aggregate C_{req} target.

9. Grid Aware Curtailment

Set of DERs n, electrical sensitivity coefficient for each DER s_i and required violation relief value I_{req} are given to the algorithm as an input and per DER curtailment allocation x_i such that $\sum_{i=1}^{n} s_i \cdot x_i \geq I_{req}$, while minimizing the total allocated curtailment $\sum_{i=1}^{n} x_i$.

Each DER is also assigned their sensitivity coefficient, representing its $x_i$contribution to relieving the specific network violation. First DERs are sorted according to their $s_i$values in descending order and greedily curtailed up to their $q_i$values until cumulative weighted impact is greater than required violation relief. Unlike previous methods , this algorithm directly aims at localized violation relief rather than meeting only the C alt requirement, which makes it a congestion location aware approach. The grid-aware curtailment methodology is presented in Table 5.

Sorting by s_i prioritizes the DERs providing the greatest violation relief per curtailed unit. According to the single-constraint linear sensitivity formulation, this methodology minimizes the total physical curtailment required to satisfy I_{req}.

Table 4. Merit order curtailment pseudo code

```
For every i ∈ D, calculate q_i ← k_i · a_i − min_i
```

```
Remove every DER satisfying q_i ≤ 0
Sort the remaining DERs by c_i in ascending order
Remaining ← C_req
x_i ← 0 for every i ∈ D
for every i ∈ D in sorted order do
   if Remaining ≤ 0, break
   x_i ← min(q_i, Remaining)
   Remaining ← Remaining − x_i
end for
if Remaining > 0, report infeasible
return x
```

Table 5. Grid-aware curtailment pseudo code

```
For every i ∈ D, calculate q_i ← k_i · a_i − min_i
Remove every DER satisfying q_i ≤ 0 or s_i ≤ 0
Sort the remaining DERs by s_i in descending order
RemainingImpact ← I_req
x_i ← 0 for every i ∈ D
for every i ∈ D in sorted order do
   if RemainingImpact ≤ 0, break
   x_i ← min(q_i, RemainingImpact / s_i)
   RemainingImpact ← RemainingImpact − s_i · x_i
end for
if RemainingImpact > 0, report infeasible
return x
```

10. Computational Requirements and Theoretical Scalability

Since active network management platforms provides control over grid networks, curtailment allocation problem must be solved in real time as well. Let (n) denote the number of eligible DERs installed to grid. Minimum DER count, merit order and grid-aware algorithms each require one sorting operation followed by a linear allocation cycle, resulting in $O(n\log_n)$ time complexity and O(n) storage. With that, etheequal percentage algorithm has O (n^2) worst-case complexity when the active DER set is scanned again, right after each saturation event sorting-based implementation using saturation breakpoints can reduce this complexity to $O(n\log_n)$.

For gridaw, the curtailment methodology: O sol parantez n alt simge taban, log, son taban, alt ogy, $O(n\log_n)$ values describe only the allocation phase. Calculation and periodic update of sensitivity coefficients constitute a separate process with its own computational requirements, depending on the network model and the selected sensitivity calculation method. Therefore, sensitivity retrieval may represent a greater practical burden in computational and data management than subsequent greedy allocation.

11. Data Preparation

To create a representative scenario for comparing these algorithms, a source-based, semi-synthetic benchmark was constructed from the IEEE 123-bus qsts (quasi-static time series) feeder released through the Open Energy Data Initiative (OEDI) [21, 22]. Source-based model provides nine active PV (photovoltaic) systems, their bus and phase connections, 50 kW (kilowatt) base nameplates, and quarter-hour PV profiles. Table 6 maps domain parameters to their corresponding origin classes and construction rules.

Then, each 50-kW PV source is scaled by a factor to create a controlled, high PV operating condition. Within the voltage regulator controls held as fixed, the initial voltage at Bus 83 phase 1 was 1.061 pu (per-unit), exceeding the 1.050 pu limit. Required aggregated curtailment, which is Calt req, is determined by applying binary search to the common proportional reduction in PV output until the voltage limit has been reached [21],[22]. Meanwhile, the grid-aware target I alt req sensitivity-weighted kW-equivalent was derived from local finite-difference voltage sensitivities [8],[21],[22].

Based on direct and derived references to the sources, the nine DER benchmarks given in Table 7 are used to compare the behavior of the four curtailment algorithms. Values of h_i and q_i were calculated from $h_i = a_i - min_i$ and $q = k_i a_i - min_i$, respectively. Since the current use case does not contain any contractual information, the curtailability value k_alt, which represents the maximum allowed curtailment percentage for a specific DER, is set to 1.

Within this DER set, the topology, PV identities, base nameplates, and profile values are directly taken from the public model. Alt i, s alt i, C alt req, I alt req, and post allocation voltages are derived from power flow calculations. Since the data does not include any contractual details, all values are set to 1, meaning all DERs can be fully curtailed if needed.

Table 6.DER benchmark parameters, their referenced values, and data origins

Distributed Energy Resource Property	**Referenced value**	**Data origin**
DER pool	Nine active PV systems at buses, 7, 29,49,55,63,68,80,87,113 [21],[22]	Direct source
a_i	145.961-149.866 kW [21],[22]	Direct source
min_i	0 kW for every PV [21],[22]	Source derived
max_i	150 kW [21],[22]	Source derived
k_i	1.0	-
s_i	0.049-1.000 [8]	Source derived
c_i	80-1000 GBP/MWh [23]	Direct source
C_{req}	530.447 kW [21],[22]	Source derived
I_{req}	194.432 kW [8], [21],[22]	Source derived

Table 7. OEDI-based DER inputs used for algorithm executions

	Distributed Energy Resource Property

Distributed Energy Resource Name	a_i(kW)	min_i(kW)	max_i(kW)	k_i	h_i(kW)	q_i(kW)	c_i (GBP/MWh)	s_i
DER-07	149.583	0	150	1.00	149.583	149.583	320	0.0493
DER-29	145.961	0	150	1.00	145.961	145.961	90	0.0943
DER-49	145.961	0	150	1.00	145.961	145.961	650	0.0948
DER-55	149.283	0	150	1.00	149.283	149.283	150	0.1504
DER-63	147.350	0	150	1.00	147.350	147.350	450	0.2092
DER-68	149.548	0	150	1.00	149.548	149.548	1,000	1.0000
DER-80	145.847	0	150	1.00	145.847	145.847	80	0.2816
DER-87	145.847	0	150	1.00	145.847	145.847	250	0.2466
DER-113	149.866	0	150	1.00	149.866	149.866	500	0.9642

12. Execution Analysis

Each algorithm mentioned in the algorithmic details section has been executed on the source-based DER set in Table 7, which forms an identical input base. Resulting curtailment allocations from four methodologies were applied to the IEEE 123-bus feeder data in OpenDSS to assess how these curtailments affect grid behavior.

Total curtailment cost metric is calculated as toplam alt limit i eşittir 1 üst limit n of, c alt i,, nokta x alt i , son toplama/ 1000 in GBP/h, algorithmic feasibility stands for satisfaction of the requirements and targets, and violation resolved determined by post allocations on bus 83 right after power flow replay on OpenDSS.

According to the execution results observed with four different curtailment algorithms in Table 8, the selection of the appropriate methodology depends heavily on operational objectives. Power flow analysis after curtailment allocations indicates that meeting the aggregate curtailment target does not always eliminate voltage violations.

Table 8. Results of algorithm executions in terms of evaluation criteria

Evaluation Criteria	Algorithm Type			
	Equal percentage	**Minimum DER count**	**Merit order**	**Grid aware**
Curtailment allocation target (kW)	C_{req} = 530.448	C_{req} = 530.448	C_{req} = 530.448	I_{req} = 194.432
Number of DERs curtailed	9	4	5	2
Total curtailment (kW)	530.448	530.448	530.448	196.099
Total curtailment cost (GBP/h)	206.39	284.57	69.54	172.82
Post-allocation at Bus 83 phase 1 (pu)	1.050	1.041	1.055	1.049

Local 1.05 pu violation resolved.	Yes	Yes	No	Yes
Algorithmic feasibility	Yes	Yes	Yes	Yes
Congestion-location aware	No	No	No	Yes
Primary objective	Proportional fairness	Minimum activated DER count	Minimum direct cost	Minimum kW for localized relief
Additional data required	None	None	c_i per DER	s_i per DER

When all DERs have identical contractual constraints, and no additional data are available, such as curtailment costs and electrical sensitivities, an equal-percentage curtailment strategy is appropriate, since it is the only methodology that applies curtailment proportionally while satisfying the fairness principle. This method curtailed all nine DERs by equally distributing curtailments by %39.91 of their available capacity and reduced bus 83 phase 1 to 1.050 pu.

Scenarios that the grid operator decides the minimum number of DERs contributes to the curtailment process, and again, no additional data is available, as in the equal percentage scenario, then the minimum DER count capacity based greedy curtailment algorithm steps forward. Thanks to this methodology, communication overhead during the application of curtailment to the field equipment was kept minimal. In the benchmark set, the curtailed set was reduced from nine to four, and it is seen that the selection of DER-68 and DER-113, which are electrically effective based on their s alt ivalues, reduced bus 83 to 1.041 pu. This result indicates that the minimum DER count strategy is not location-aware and does not provide localized relief, since the DERs with the largest q alt values are also sensitive assets.

During curtailment operations, if cost efficiency is the operator's top priority, the merit order curtailment algorithm is used, as its primary objective is to keep the cost of the curtailment process to a minimum. With *values set to 1, the algorithm curtailed five DERs and produced the lowest total curtailment cost of 69.54* GBP. Post-allocation voltage remained at 1.055 pu, exceeding the 1.05 pu limit, since the allocation order does not consider DERs' voltage sensitivities with respect to the violated bus. Also, the curtailment cost of each DER should be determined within the grid to benefit from this method.

When a violation is observed on a specific line or on overloaded equipment on the grid, the grid-aware curtailment algorithm is the only option aimed at localized overload relief. With k_i values set to 1, the algorithm selected DER-68 and DER-113, which required 196.099 kW of total curtailment and reduced us 83 to 1.049 pu. The total curtailment value of this method must not be directly compared with other approaches, since it satisfies I alt req rather than C alt req while meeting the .050 pu voltage limit. Also, this methodology requires precomputed sensitivity values for each DER within the fleet.

Minimum DER count, merit order, and grid-aware allocation strategies require one sorting operation followed by a linear pass; therefore, they all have O (n log n) time complexity. Naïve iterative water filling implementation of equal percentage cut-off has $O(n^2)$ worst-case complexity w,h wwhereas ssortingthe DDERS aaccording to theirsaturation thresholds reduces tthe complexityto Oo OOO complexity to OOO complexity to OOO complexitytoO$(n\log_n)$.



IV. CONCLUSION & FUTURE WORK

This study compared four deterministic DER curtailment allocation strategies representing distinct operational objectives: proportional burden sharing, minimum DER activation, operational cost minimization, and location-specific violation relief. These methodologies were evaluated using a source-based semi-synthetic benchmark derived from the IEEE 123-bus QSTS feeder and nine PV profiles available in the Open Energy Data Initiative. The resulting allocations were observed in OpenDSS, examining both the algorithmic feasibility of each strategy and its actual effects on the grid. Experimental analysis indicates that no single algorithm outperforms the others when considering all objectives included in this study. The choice of methodology depends on the grid operator's operational priorities. Data availability, such as curtailment cost or electrical sensitivity data for each DER, also affects which algorithm to choose when calculating curtailment. The most immediate next step is to apply all four curtailment algorithms to real operational DER datasets and reproduce the same comparative analysis on data retrieved from actual grid networks.

Increasing algorithm diversity also represents a promising direction for future research. New methodologies, such as LP-based optimal power flow formulations and hybrid greedy optimization approaches, may also prove useful for congestion resolution in various aspects.

REFERENCES

[1] R. A. F. Currie, G. W. Ault, R. W. Fordyce, D. F. MacLeman, M. Smith, and J. R. McDonald, "Actively Managing Wind Farm Power Output," *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1523–1524, Aug. 2008, doi: 10.1109/TPWRS.2008.926722.

[2] R. Tonkoski, L. A. C. Lopes, and T. H. M. El-Fouly, "Coordinated Active Power Curtailment of Grid Connected PV Inverters for Overvoltage Prevention," *IEEE Transactions on Sustainable Energy*, vol. 2, no. 2, pp. 139–147, Apr. 2011, doi: 10.1109/TSTE.2010.2098483.

[3] M. Moness and A. M. Moustafa, "A Survey of Cyber-Physical Advances and Challenges of Wind Energy Conversion Systems: Prospects for Internet of Energy," *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 134–145, Apr. 2016, doi: 10.1109/JIOT.2015.2478381.

[4] R. J. Hennig, L. J. de Vries, and S. H. Tindemans, "Congestion management in electricity distribution networks: Smart tariffs, local markets and direct control," *Utilities Policy*, vol. 85, p. 101660, Dec. 2023, doi: 10.1016/j.jup.2023.101660.

[5] F. García-Muñoz, A. Ivanova, J. Farré Montané, M. Serrano, and C. Corchero, "A DSO flexibility platform based on an optimal congestion management model for the European CoordiNet project," *Electric Power Systems Research*, vol. 221, p. 109386, Aug. 2023, doi: 10.1016/j.epsr.2023.109386.

[6] B. Stott, J. Jardim, and O. Alsac, "DC Power Flow Revisited," *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1290–1300, Aug. 2009, doi: 10.1109/TPWRS.2009.2021235.

[7] S. Yoshizawa *et al.*, "Voltage-Sensitivity-Based Volt-VAR-Watt Settings of Smart Inverters for Mitigating Voltage Rise in Distribution Systems," *IEEE Open Access Journal of Power and Energy*, vol. 8, pp. 584–595, 2021, doi: 10.1109/OAJPE.2021.3125013.

[8] K. Knez, L. Herman, and B. Blažič, "Dynamic Management of Flexibility in Distribution Networks through Sensitivity Coefficients," *Energies*, vol. 17, no. 7, p. 1783, Jan. 2024, doi: 10.3390/en17071783.

[9] F. Capitanescu *et al.*, "State-of-the-art, challenges, and future trends in security constrained optimal power flow," *Electric Power Systems Research*, vol. 81, no. 8, pp. 1731–1741, Aug. 2011, doi: 10.1016/j.epsr.2011.04.003.

[10] D. Gebbran, S. Mhanna, Y. Ma, A. C. Chapman, and G. Verbič, "Fair coordination of distributed energy resources with Volt-Var control and PV curtailment," *Applied Energy*, vol. 286, p. 116546, Mar. 2021,

doi: 10.1016/j.apenergy.2021.116546.
[11] S. Poudel, M. Mukherjee, R. Sadnan, and A. P. Reiman, "Fairness-Aware Distributed Energy Coordination for Voltage Regulation in Power Distribution Systems," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 3, pp. 1866–1880, Jul. 2023, doi: 10.1109/TSTE.2023.3252944.
[12] "Operation of power systems," *Wind Power Integration: Connection and System Operational Aspects*, pp. 137–216, doi: 10.1049/PBRN014E_ch5.
[13] D. Pudjianto, C. Ramsay, and G. Strbac, "Virtual power plant and system integration of distributed energy resources," *IET Renewable Power Generation*, vol. 1, no. 1, pp. 10–16, Mar. 2007, doi: 10.1049/iet-rpg:20060023.
[14] L. F. Ochoa and G. P. Harrison, "Minimizing Energy Losses: Optimal Accommodation and Smart Operation of Renewable Distributed Generation," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 198–205, Feb. 2011, doi: 10.1109/TPWRS.2010.2049036.
[15] P. Olivella-Rosell *et al.*, "Optimization problem for meeting distribution system operator requests in local flexibility markets with distributed energy resources," *Applied Energy*, vol. 210, pp. 881–895, Jan. 2018, doi: 10.1016/j.apenergy.2017.08.136.
[16] "Discrete-Variable Extremum Problems: https://www.jstor.org/stable/167356
[17] M. Menghwar, J. Yan, Y. Chi, M. Asim Amin, and Y. Liu, "A market-based real-time algorithm for congestion alleviation incorporating EV demand response in active distribution networks," *Applied Energy*, vol. 356, p. 122426, Feb. 2024, doi: 10.1016/j.apenergy.2023.122426.
[18] M. E. Baran and F. F. Wu, "Optimal capacitor placement on radial distribution systems," *IEEE Transactions on Power Delivery*, vol. 4, no. 1, pp. 725–734, Jan. 1989, doi: 10.1109/61.19265.
[19] E. Dall'Anese, H. Zhu, and G. B. Giannakis, "Distributed Optimal Power Flow for Smart Microgrids," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1464–1475, Sep. 2013, doi: 10.1109/TSG.2013.2248175.
[20] Z. Wei, F. De Nijs, J. Li, and H. Wang, "Model-Free Approach to Fair Solar PV Curtailment Using Reinforcement Learning," in *Proceedings of the 14th ACM International Conference on Future Energy Systems*, in e-Energy '23. New York, NY, USA: Association for Computing Machinery, Haziran 2023, pp. 14–21. doi: 10.1145/3575813.3576871.
[21] T. Elgindy and K. Balasubramaniam, "Sample IEEE123 Bus system for OEDI SI." DOE Open Energy Data Initiative (OEDI); National Renewable Energy Laboratory, p. 4 files, 2022. doi: 10.25984/2228282.
[22] *openEDI/oedisi-ieee123*. (Apr. 26, 2026). Open Energy Data Initiative. Accessed: Aug. 02, 2026. [Online]. Available: https://github.com/openEDI/oedisi-ieee123
[23] SP Energy Networks, "Distribution Flexibility Service: Participation Guidance," Jul. 2025. Accessed: Aug. 2, 2026. [Online]. Available: https://www.spenergynetworks.co.uk/userfiles/file/SPEN_Month_Ahead_Market_Participation_Guidance_2025.pdf

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

TRITON: A Modular Transformer for Per-Pixel Marine Pollutant Detection in Sentinel-2 Imagery

José Luis de León Carmona
Center for Research in Mathematics (CIMAT)
Guanajuato, Mexico
jose.deleon@cimat.mx

Rogelio Hasimoto-Beltrán
Center for Research in Mathematics (CIMAT)
Guanajuato, México
hasimoto@cimat.mx

Abstract
We present a per-pixel classification system for marine surface monitoring in Sentinel-2 multispectral imagery. The method rests on two contributions. First, *pixel2vects*, a feature-engineering process that converts an unstructured multispectral hypercube into a structured dataset by precomputing a 68-dimensional vector per pixel 11 raw spectral bands, 12 physics-based spectral indices, and 45 spatial descriptors (multiscale, texture, and band-context). This makes local spatial context explicit and the sparse-annotation problem of remote sensing is turned into a dense tabular one. Second, TRITON (**Token Representation of Imagery for Transformer Orchestration Networks**), a 3.35M-parameter modular Transformer that treats each of the five heterogeneous feature groups as an independent token and models their interaction through self-attention. To handle the severe class imbalance in the MADOS benchmark (a 405× ratio across 15 classes), TRITON is trained with a Label-Distribution-Aware Margin (LDAM) loss combined with a prototype-memory auxiliary loss. On the MADOS test set, our best single-seed model reaches 76.43% macro-F_1, surpassing the published state of the art, MariNeXt (76.0%), without patch-level CNNs, pretraining, or copy-paste augmentation. This represents a more efficient solution for marine pollution detection in both training and inference.

Keywords: marine pollutant detection, oil spill, Sentinel-2, Transformer, class imbalance, per-pixel classification, MADOS.

I. INTRODUCTION

Detecting marine pollutants—oil spills, floating debris, sea snot, and related surface phenomena—from satellite imagery is a prerequisite for large-scale, continuous ocean monitoring. Sentinel-2 offers a favorable combination of spatial resolution (10–20 m), spectral coverage (11 usable bands between 440 nm and 2200 nm), and a 5-day global revisit cycle. The MADOS dataset proposed in [1] formalizes the problem as a 15-class per-pixel classification over 174 scenes, partitioned at the scene level to prevent spatial leakage between splits. Recent works on the MADOS dataset have largely relaxed one of its two hard constraints—either the labeling requirement, through self-supervised transfer to other sensors [2], or the class structure itself, by collapsing the 15 classes into a debris-versus-background decision that makes the severe imbalance of MADOS more tractable [3]. We keep both constraints and instead change the pixel representation. A central difficulty with remote-sensing datasets is sparse annotation: of the 2803 MADOS patches (240 × 240 pixels), only 1,481,155 (0.91%) pixels are labeled; the rest are marked as unlabeled. Approaches that take in the multispectral image, directly translate this sparsity into heavy CNN or segmentation architectures with large number of parameters and high training and inference cost. This work takes a per-pixel view instead, hypothesizing that the spectral signature of a pixel, along with its local spatial context, contains most of the information needed to classify it, which is under-exploited by scene-level models. We therefore precompute, for every pixel, a fixed-length vector that encodes both spectral and spatial evidence (Section IV) and classify each vector independently using a compact Transformer (Section V). The formulation has practical

advantages: predictions are independent of scene size, computation can be trivially parallelized, the sparse-annotation problem becomes a dense tabular one with millions of labeled examples, and every input feature has an explicit physical definition. The contributions of this work are: pixel2vects, a 68-feature per-pixel representation for Sentinel-2 marine scenes, organized into five physically meaningful groups (bands, indices, multiscale, texture, band-context), whose spatial component accounts for the largest single performance gain we observe (≈+11 pts macro-$F1$). TRITON, a 3.35M-parameter modular Transformer that tokenizes feature groups rather than individual features, projecting each group into a shared embedding space and modeling inter-group interaction with self-attention. An imbalance-handling recipe balanced sampling with an LDAM margin loss and a prototype-memory auxiliary term together with an ablation showing which additional mechanisms (Deferred Re-Weighting, post-hoc logit adjustment, cross-loss ensemble) are counterproductive once balanced sampling is in use. State-of-the-art results on MADOS together with a reproducible multi-seed evaluation protocol.

II. RELATED WORK

Kikaki et al. [1] introduced MADOS dataset and evaluated a spectrum of baselines. Among per-pixel methods, Random Forest [4] over spectral features reaches 56.6% macro-$F1$, and RF++—the same classifier augmented with hand-crafted spatial features—reaches 64.4%. Among scene-level segmentation methods, U-Net [5] obtains 63.8%, SegNeXt [6] 60.6%, and MariNeXt, a SegNeXt-based pipeline with multi-scale features, Very Simple Copy-Paste (VSCP) augmentation, and geometric test-time augmentation, holds the published state of the art at 76.0%. The 7.8-pt gap between RF and RF++ is direct evidence that explicit spatial descriptors carry substantial information in this domain; our pixel2vects representation extends this observation. Recent work in this area has moved in three directions. Dalsasso et al. [2] attack the labeling bottleneck rather than the model, transferring debris detection from Sentinel-2 to daily 3-m PlanetScope acquisitions through self-supervised cross-sensor pre-training—trading spectral richness for revisit frequency. Bouchelaghem et al. [3] merge MARIDA and MADOS and collapse the label space to debris versus background, reporting that models trained on MADOS transfer to MARIDA with a relative $F1$ loss of only 1.25% (0.901 → 0.890), which suggests that the geographic diversity of the benchmark matters more for generalization than its annotation density. Giannopoulos et al. [7] used ConvNeXt screening stage with a domain-adapted Segment Anything Model for oil spills in Sentinel-1 SAR and—directly relevant to our design—find it necessary to append GLCM texture measures to the backscatter input, an independent instance of hand-crafted spatial descriptors improving a modern deep model. None of the three retains the 15-class per-pixel formulation this work addresses. Focal loss [8] re-weights the per-example gradient by prediction difficulty. LDAM [9] instead enforces a per-class margin $\Delta y \propto n{-}1/4$ on the decision boundary, optionally combined with Deferred Re-Weighting (DRW), a two-phase schedule that introduces inverse-frequency class weights late in training. Logit adjustment [10] corrects model outputs post hoc using the training label prior. A blunter alternative is to remove the imbalance by construction, collapsing the label space into a binary decision [3]; we retain all 15 classes and treat the imbalance as a loss-geometry problem. These mechanisms are typically evaluated in isolation; Section VI examines their behavior in combination with balanced re-sampling. Vaswani et al. [11], introduced a multi-head self-attention Transformer (originally for language processing) over a token sequence, while FT-Transformer [12] adapts it to tabular data by mapping each scalar feature to a token and reading out through a CLS (classification) token. In remote sensing the prevailing per-pixel Transformer design instead attends over a spectral–spatial neighborhood: Wu et al. [13] classify hyperspectral pixels by combining a convolutional front-end with cross-window attention over the patch surrounding each pixel, so spatial context enters the model as attended raw values. TRITON adopts the token-plus-CLS scheme but tokenizes at the granularity of physically defined feature groups (Section V), which shortens the sequence from 69 to 6 tokens and respects the heterogeneous numeric regimes of the groups; spatial context enters as precomputed

descriptors rather than as an attended neighborhood, which is what keeps the sequence short and each input interpretable.

III. DATASET

MADOS [1] comprises 174 Sentinel-2 scenes acquired between 2015 and 2022, annotated by domain experts into 15 marine surface classes: Marine Debris, Dense Sargassum, Sparse Floating Algae, Natural Organic Material, Ship, Oil Spill, Marine Water, Sediment-Laden Water, Foam, Turbid Water, Shallow Water, Waves & Wakes, Oil Platform, Jellyfish, and Sea Snot. The split is performed at the scene level (≈70/15/15), so no geographic region contributes pixels to more than one partition. We use all 11 non-atmospheric bands (B1–B8A, B11, B12; 440–2200 nm) resampled to 10 m, with ACOLITE Rayleigh-corrected reflectance (ρrc) following the MADOS preprocessing protocol. After feature extraction (Section IV), our working dataset contains 1,078,367 training, 203,652 validation, and 199,136 test pixels. Two properties of the benchmark shape everything that follows. First, annotation is sparse: of the 2,803 patches (240×240 pixels), only 0.91% of pixels carry a label. Second, the training distribution is severely long-tailed—Marine Water contributes on the order of 105 pixels against 1,165 for Natural Organic Material and 1,339 for Jellyfish. Because macro-$F1$ weights every class equally, performance on the rare classes dominates the metric, which is what drives the loss design of Section V and the ablations of Section VI.

IV. PIXEL2VECTS: PER-PIXEL FEATURE REPRESENTATION

A Sentinel-2 multispectral image is defined as $I \in \mathrm{R}H\times W\times 11$. The pixel2vects process, maps every pixel into a 68-dimensional feature vector

$$\mathbf{f} = [\mathbf{b}, \mathbf{i}, \mathbf{m}, \mathbf{t}, \mathbf{c}], \tag{1}$$

that is, it concatenates five module vectors—bands, indices, multiscale, texture and band-context. The grouping is not cosmetic: it defines the token structure of TRITON (Section V) and the granularity at which normalization statistics are computed, each group being independently z-score normalized with training-set statistics. Bands (11): Raw ρrc reflectance at the 11 non-atmospheric Sentinel-2 MSI bands: coastal aerosol (B1, 443 nm) at 60 m; blue, green, red and NIR (B2–B4, B8; 490, 560, 665, 842 nm) at 10 m; and the three red-edge bands (B5–B7; 705, 740, 783 nm), the narrow NIR band (B8A, 865 nm) and the two SWIR bands (B11, 1610 nm; B12, 2190 nm) at 20 m. The two atmospheric bands, B9 (water vapor) and B10 (cirrus), carry no surface signal and are discarded. Indices (12): Physics-based band combinations that amplify class-discriminative contrast, each following its standard published definition. They cover floating vegetation (NDVI, EVI, FAI), floating debris (FDI, NRD), open and turbid water (NDWI, MNDWI, AWEI, SI), moisture and surface films (NDMI), and sediment or bare substrate (BSI, NDSII). Six belong to the normalized-difference family $(bi - bj)/(bi + bj)$, while FAI and FDI are baseline-subtraction indices measuring the NIR excess of a pixel over a red–SWIR linear baseline—the standard signature of floating material.

Multiscale (20): Computed on the gray scale composite $I_{gray} = (B2 + B3 + B4)/3$ at five scales $\sigma \in \{1, 2, 4, 8, 16\}$ pixels (i.e., 10-160 m): the Gaussian-smoothed intensity G_σ, the Sobel gradient magnitude $M_\sigma = \sqrt{(S_\sigma^x)^2 + \left(S_\sigma^y\right)^2}$, and the two eigenvalues $(\lambda_\sigma^1 + \lambda_\sigma^2)$ of the local Hessian of G_σ, 5 + 5 + 10 = 20 features. The eigenvalue pair characterizes local geometry—blob-like structures ($\lambda 1 \approx \lambda 2 < 0$) versus elongated ridges ($|\lambda 1| \gg |\lambda 2|$)—which helps separate slick patches from ship wakes. Texture (13): Six

GLCM/Haralick descriptors (contrast, homogeneity, dissimilarity, ASM, energy, correlation) from the per-pixel Grey-Level Co-occurrence Matrix over a 13 × 13 window with 16 gray levels, accumulated over four directions (0◦, 45◦, 90◦, 135◦) [14]; standard and uniform Local Binary Patterns at $P = 8$, $R = 1$, each normalized to [0, 1] (2) [15]; the Laplacian of Gaussian $\nabla 2G\sigma$ at $\sigma \in \{2, 4\}$ (2) [16]; and the local standard deviation over windows $w \in \{5, 11, 21\}$, computed via $Var[X] = E[X^2] - E[X]^2$ (3). Band-context (12): Local descriptors over the bands most informative for floating-material detection (B4, B8, B11) and NDVI: the 5 × 5 local mean and standard deviation of each of B4, B8 and B11 (6) and of NDVI (2); the local gradient direction at $\sigma = 1$ encoded as $(cos\theta, sin\theta)$ to avoid the $\pm\pi$ discontinuity (2); and the two cross-band local mean differences Δ_{B8-B4} and Δ_{B8-B11} (2). The 45 spatial descriptors of the last three groups are computed once per patch with vectorized operations and stored in HDF5, so the per-pixel classifier receives spatial context as a precomputed input.

V. TRITON

TRITON follows the encoder-Transformer design of Vaswani et al. [11]. The 68-feature vector is partitioned into its five module vectors of dimension $d$bands = 11, $d$idx = 12, $d$ms = 20, $d$tex = 13, $d$bctx = 12. Because these groups occupy heterogeneous numeric and physical regimes (reflectances ≈0.01–0.3, GLCM contrast in [0, 10], LBP codes in [0, 255]), TRITON projects each group independently before letting them interact through self-attention. Figure 1 summarizes the model.

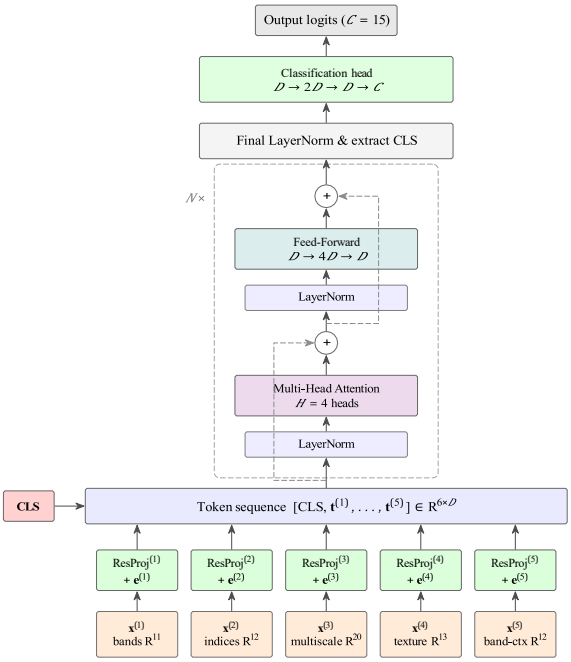


Figure 1. Architecture of TRITON. Each per-pixel input is split into five heterogeneous feature groups x(m) (68 features total); each group is independently projected to a shared D-dimensional space by a ResidualProjection block and augmented with a learnable module-type embedding e(m) . With a prepended learnable CLS token, the resulting 6-token sequence is processed by $N = 3$ pre-norm Transformer encoder layers with residual connections, and the final CLS embedding is mapped to $C = 15$ class logits by a three-layer head.

From vectors to tokens: Each module vector x(m) ∈ Rdm is mapped to a token of shared dimension $d$model = 256 through a residual projection:

$$\mathbf{z}^{(m)} = \mathrm{LN}_{\mathrm{out}}(\,\mathrm{MLP}^{(m)}(\mathrm{LN}_{\mathrm{in}}(\mathbf{x}^{(m)})\,) + \mathbf{W}^{(m)}_{\mathrm{skip}}\,\mathrm{LN}_{\mathrm{in}}(\mathbf{x}(m)), \tag{2}$$

with MLP(m) (h) = W2 Dropout0.20(GELU(W1h + b1)) + b2 and FFN hidden width 2 $d$model. The linear skip path and the output LayerNorm stabilize optimization across groups with disparate input statistics.

A learnable module-type embedding e(m) ∈ R$d$model —the analogue of positional encodings for an unordered set of typed tokens—is added to each token, and a learnable class token is prepended, giving the 6-token sequence T (0) = [$t$cls, t (1) , . . . , t (5)] with t (m) = z(m) + e(m).

Encoder and classification head: The sequence passes through L = 3 pre-norm Transformer encoder layers (h = 4 attention heads, head dimension 64, FFN ratio r = 4 so $d$ff = 1024, attention dropout 0.05, FFN dropout 0.10), each applying multi-head self-attention and a feed-forward block with residual connections around both, followed by a final LayerNorm. The class-token output hcls aggregates information from all module tokens and feeds a 3-layer head (256 → 512 → 256 → 15, GELU, dropout 0.20) producing class logits ℓ ∈ R15; at inference yˆ = arg max$c$ ℓc. The complete model has 3,352,215 parameters.

Loss function: The training objective combines a margin loss on the classifier logits with a prototype loss on the embedding geometry:

$$\mathrm{L} = \mathrm{L}_{\mathrm{LDAM}}(\ell,\ y) + \lambda_{\mathrm{proto}} \cdot \mathrm{L}_{\mathrm{proto}}(\ell_{\mathrm{proto}},\ y), \tag{3}$$

with λ_{proto} = 0.3.

LDAM: To address class imbalance [9], a per-class margin is subtracted from the true-class logit during training:

$$\Delta_c = \min\left(\frac{C}{n_c^{1/4}},\ \Delta_{\max}\right) \qquad \tilde{\ell}_y = \ell_y - \Delta_y, \tag{4}$$

(other logits unchanged), and standard cross-entropy is applied to ℓ˜. Rare classes are thus required to be classified with a wider safety margin. We use C = 0.5, Δmax = 0.5, and disable label smoothing (which interacts destructively with the margin geometry).

Prototype memory: An exponential-moving-average bank maintains one unit-norm centroid $\mathbf{p}_c$ per class in CLS space, updated each batch from the normalized class-mean embedding with momentum α = 0.999. Auxiliary logits are cosine similarities scaled by temperature τ = 0.07,

$$\ell_{\text{proto, }\square} = \frac{1}{\square} \cdot \frac{\langle \mathbf{h}^{\text{cls}}, \mathbf{p}_{\square} \rangle}{\| \mathbf{h}^{\text{cls}} \|}, \tag{5}$$

and the auxiliary cross-entropy encourages compact, well-separated class clusters on the unit sphere—targeting spectrally confusable pairs such as Oil Spill / Marine Water. The two terms act on complementary spaces: LDAM shapes the decision boundary in logit space, the prototype loss structures the embedding geometry.

VI. EXPERIMENTS

All experiments use the MADOS scene-level splits; macro-$F1$ on the test set is the primary metric, with mIoU and overall accuracy (OA) as secondary metrics. TRITON is trained with WeightedRandomSampler (exponent β = 0.7, sampling probability $\propto n-\beta$), the LDAM + prototype loss of Section V, Gaussian input noise (σ = 0.02) on bands and indices, band dropout (p = 0.15), mixup (α = 0.1) applied jointly to all five groups, gradient clipping at norm 1.0, AdamW (lr 3 × 10−4, weight decay 10−4) and OneCycleLR with 10% warm-up; one run takes 35–80 minutes on a single NVIDIA GPU. Every epoch's checkpoint is scored on validation macro-$F1$, the top-5 are retained and their logits averaged at evaluation, and test-time augmentation (TTA) averages logits over 8 forward passes with independent Gaussian noise (σ = 0.02) on the spectral features.

1. Results on MADOS

Table 1. MADOS test-set results. Best result per metric in bold. All values in %.

Method	Macro-F_1	mIoU	OA
RF [1]	56.6	43.9	67.1
RF++ [1]	64.4	52.4	83.8
U-Net [5]	63.8	51.0	82.9
SegNeXt [6]	60.6	49.2	86.6
MariNeXt (pub. SOTA)	76.0	64.3	**89.1**
TRITON – best single (seed 123)	**76.43**	**66.55**	88.69
TRITON – 3-seed mean ± std	75.72±0.68	—	—
TRITON – 15-ckpt ens. × TTA-8	76.18	66.38	88.73

As shown in Table 1, TRITON exceeds the published state of the art in macro-$F1$ (+0.43 best single, +0.18 in

the 15-checkpoint ensemble—hereafter "ensemble") and mIoU (+2.25 / +2.08), with OA 0.4 pts lower—the expected consequence of reallocating decision margin from frequent to rare classes under a metric that values the latter. Table 2 breaks the deployed system (15-checkpoint ensemble × TTA-8) down over the 199,136 annotated test pixels. Oil Spill, the class of primary applied interest, is the strongest pollutant class at 81.8% $F1$, and its error is asymmetric in the operationally favorable direction: 89.2% recall against 75.5% precision, i.e., the system misses one in nine true slick pixels while over-flagging at a rate an analyst can triage. Nine of the fifteen classes exceed 85% $F1$ and Sediment-Laden Water is effectively solved (99.7%).

Three classes carry the macro-average penalty. Natural Organic Material collapses to 6.8% $F1$, but with 31 test pixels any estimate is statistically meaningless—a data-scarcity problem (1,165 training pixels) rather than a modeling failure. Waves & Wakes (44.6%) and Jellyfish (45.4%) are genuinely hard: both keep reasonable precision (65.4% and 81.9%) but Recall metric below 34%, so the models are conservative on them rather than confused. Relative to the best single model the ensemble redistributes rather than adds accuracy: Oil Spill improves from 75.2 to 81.8% $F1$ while Jellyfish (78.0 → 45.4) and Waves & Wakes (65.7 → 44.6) give back the margin LDAM had recovered on individual seeds, since averaging logits over 15 checkpoints pulls the rare-class boundaries back toward the frequent classes they border. The two systems reach nearly the same macro-$F1$ (76.43 vs. 76.18%) by opposite routes: the ensemble here buys independence from seed choice, not tail performance.

Table 2. Per-class results of the deployed system (15-checkpoint ensemble × TTA-8). Support in pixels; all other values in %.

Class	Support	Prec.	Rec.	F_1	IoU
Marine Debris	836	72.1	83.4	77.3	63.0
Dense Sargassum	472	94.6	85.4	89.8	81.4
Sparse Floating Algae	982	89.5	90.9	90.2	82.2
Natural Organic Mat.	31	7.1	6.5	6.8	3.5
Ship	2,048	78.6	72.3	75.3	60.4
Oil Spill	38,488	75.5	89.2	81.8	69.2
Marine Water	103,394	94.5	87.8	91.0	83.5
Sediment-Laden Water	25,038	99.4	100.0	99.7	99.4
Foam	223	78.6	98.7	87.5	77.7
Turbid Water	8,557	76.9	96.1	85.4	74.5
Shallow Water	10,621	87.1	83.9	85.4	74.6
Waves & Wakes	3,324	65.4	33.9	44.6	28.7
Oil Platform	2,614	83.1	93.4	87.9	78.5
Jellyfish	274	81.9	31.4	45.4	29.4
Sea Snot	2,234	97.8	91.6	94.6	89.7
Macro avg	199,136	78.8	76.3	**76.18**	**66.38**
Overall accuracy					88.73

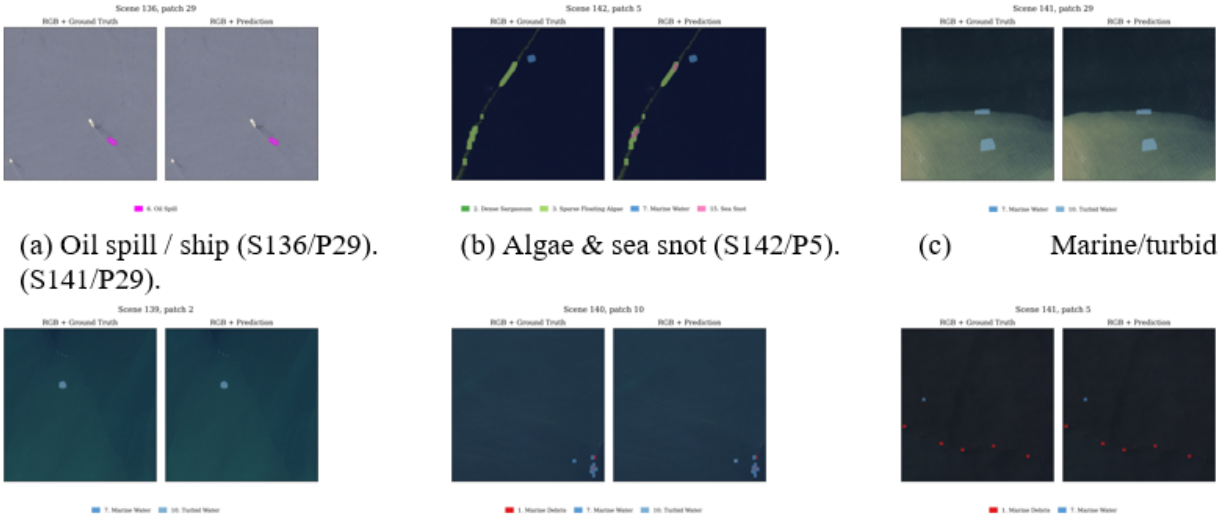
Table 3. Row-normalized confusion matrix of the deployed system, in % of the row support (rows: ground truth; columns: predictions, by the class index of the first column; bold diagonal: recall). "·" is nonzero but below 0.5%, blank is exactly zero.

		Predicted class													
#	True class	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

0	Marine Debris	**83**		2	·	12			1		·				·	·	·	2				
1	Dense Sargassum		**85**	12																		2
2	Sparse Floating Algae		2	**91**				4							2	·						1
3	Natural Organic Mat.	55		13	**6**	23										3						
4	Ship	1			·	**72**		4	·		·	·	·			1	20					·
5	Oil Spill	·		·		·		**89**	10			·	·			1	·					
6	Marine Water	·		·		·	9		**88**		·	1	1			·						·
7	Sediment-Laden Water									**100**												
8	Foam										**99**					1						
9	Turbid Water								3	1	**96**					·						
10	Shallow Water			·		1			6		·	9	**84**			·						
11	Waves & Wakes	4				4	32		15		1	1	8			**34**	1			·	·	
12	Oil Platform					6	1									·	**93**		·			
13	Jellyfish	5					2	57										3	**31**	1		
14	Sea Snot	3			·	1	·	·	1		3						·				**92**	

Table 3 shows that the errors are structured, not diffuse. Most of the off-diagonal mass falls in a single dark-surface cluster spanning Oil Spill, Waves & Wakes, Jellyfish and Marine Water: 32% of Waves & Wakes pixels and 57% of Jellyfish pixels are predicted as Oil Spill, and together with the 9% of Marine Water assigned to Oil Spill they account for essentially all of that class's precision loss, while the reverse leak is smaller and one-directional (10% of Oil Spill becomes Marine Water). Two further entries are equally interpretable: 20% of Ship becomes Oil Platform—both are bright man-made structures on water, distinguished mainly by motion, which a single acquisition cannot show—and 12% of Dense Sargassum becomes Sparse Floating Algae, a density threshold on the same material rather than a different material. The residual mass follows the same logic among adjacent water states (Shallow Water: 9% → Turbid, 6% → Marine). Errors therefore track physical similarity between surface types rather than random noise, and none place a pollutant class inside an unrelated category.

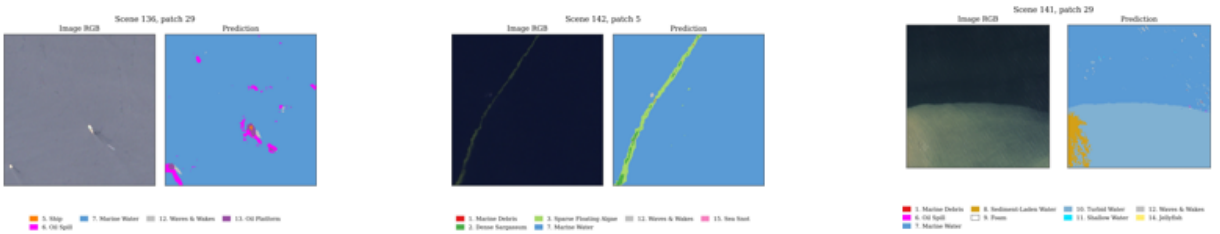
The same behavior is visible directly in the predicted maps. Figure 2 shows six representative test patches spanning the pollutant-relevant classes of the benchmark. For each patch the left panel overlays the sparse expert annotation on the RGB composite and the right panel overlays the TRITON prediction on the same composite; only annotated pixels are colored, so the two are directly comparable. The prediction tracks the ground truth closely: the Oil Spill trailing the ship in Figure 2a is recovered with its correct extent and position, the floating-vegetation trail in Figure 2b is likewise captured, and the Marine / Turbid water boundary (Figures 2c and 2d) is delineated cleanly despite the two classes being spectrally similar. The small Marine Debris targets in Figures 2e and 2f—each a handful of pixels—are localized correctly, consistent with the per-pixel formulation's ability to classify isolated targets independently of scene context. The most visible failure is



(a) Oil spill / ship (S136/P29). (b) Algae & sea snot (S142/P5). (c) Marine/turbid (S141/P29).

(d) Turbid patch (S139/P2). (e) Debris cluster (S140/P10). (f) Debris targets (S141/P5).

Figure 2. Qualitative results on six MADOS test patches. Per patch, the left panel is the RGB composite with the ground-truth annotation overlaid and the right panel the same composite with the TRITON prediction; only expert-annotated pixels are colored. Scenes are identified as S{scene}/P{patch} following the MADOS naming.



(a) Oil spill / ship (S136/P29). (b) Algae & sea snot (S142/P5). (c) Marine/turbid (S141/P29).

Figure 3. Dense per-pixel prediction maps for three of the patches of Figure 2. For each patch, the left panel is the RGB composite and the right panel is the full TRITON class map over every pixel (not only the annotated ones). The legend lists all classes appearing in the map. in Figure 2b, where a Sparse Floating Algae streak is partially predicted as Sea Snot: the floating-material confusion of Table 3, a genuine spectral overlap rather than a localization error.

Because MADOS annotation is sparse, Figure 2 only reveals behavior on labeled pixels; Figure 3 shows the full per-pixel prediction maps for the same patches. Despite classifying each pixel independently, TRITON produces spatially coherent regions—contiguous water bodies, connected vegetation trails and compact debris clusters—rather than salt-and-pepper noise, and its predictions on unlabeled pixels (e.g., the extended Oil Spill and Sea Snot regions around the annotated targets) remain consistent with the surrounding surface context.

This is the clearest visual evidence that the precomputed spatial descriptors, not scene-level context, carry the information the model needs.

2. Ablations

Representation. Comparing spectral-only per-pixel Transformers on the 11 raw bands against TRITON on the full 68-feature representation isolates the contribution of pixel2vects (Table 4, top). Spectral-only architectures plateau near 61%, below the RF++ spatial baseline (64.4%); varying architecture and schedule across designs moved test macro-$F1$ within a ≈10-pt band, all of it below that baseline. Adding the 57 spatial and index features raises TRITON to 72.35% (+11.5 pts), the largest single gain in the study attributable to the representation, not to architectural novelty. Tripling the parameters (9.26M) raises validation $F1$ by 3.5 pts but leaves test $F1$ unchanged while doubling the validation–test gap: capacity is not the bottleneck on this dataset.

Loss geometry: With the representation fixed, replacing Focal loss by the LDAM margin loss of Section V improves macro-$F1$ by +3.58 pts, with a clear optimum at $C = 0.5$ (at the end of Table 4). The gain decomposes almost entirely into the rare classes (Focal → LDAM, per-class $F1$): Jellyfish 32.9 → 80.0 (+47.1), Foam 39.4 → 78.3 (+38.9), Waves & Wakes

Table 4. Ablations on the representation (top) and on the loss geometry with the representation fixed (bottom). Test macro-F_1 in %.

Configuration	Params	Test F_1
Spectral Transformer (11 bands, wavelength PE)	1.97M	58.76
FT-Transformer (11 bands, per-feature tokens)	2.72M	60.87
TRITON (68 features, group tokens)	3.35M	**72.35**
TRITON (68 feat., D=384, L=4)	9.26M	72.51
Focal (γ = 1.5)	3.35M	72.35
LDAM (C = 0.3)	3.35M	75.46
LDAM (C = 0.5)	3.35M	**75.93**
LDAM (C = 0.7)	3.35M	75.48

49.9 → 65.7 (+15.8), Marine Debris 58.4 → 71.5 (+13.1), Sparse Floating Algae 78.6 → 88.6 (+10.0), while no frequent class loses more than 3.5 pts. This asymmetry—double-digit gains on rare classes against low single-digit losses on frequent ones—is the signature of a margin reallocation, and it is what macro-$F1$ rewards. Imbalance mechanisms do not compose: Three individually reasonable mechanisms degrade performance once balanced sampling and LDAM margins are in place. Deferred Re-Weighting (inverse-frequency class weights from epoch 50) cost

≈11 pts of validation macro-$F1$ at the switching epoch and ended at 75.62%, below plain LDAM: it interacts destructively with a sampler that already deforms the class distribution. Post-hoc logit adjustment, $\ell y \leftarrow \ell y - \tau \log \pi y$, assumes training under the prior natural label, and then degrades monotonically—72.35% ($\tau = 0$) →

66.09% (τ = 0.2) → 35.17% (τ = 1)—because the sampler has partially uniformized that prior label. Cross-loss ensemble of Focal and LDAM checkpoints yields 75.74%, below LDAM alone: the Focal models' near-zero confidence on Foam and Jellyfish dilutes precisely what LDAM repaired (Jellyfish 80% → 57% within the ensemble). Re-sampling, re-weighting, margin shaping and posterior correction all act on the same class prior, so their effects compound rather than add; one intervention point should be chosen and the rest held neutral.

Seed variance and ensemble: Across three seeds (Table 5) the spread is 1.62 pts, which exceeds the margin separating our best model from MariNeXt (+0.43 pts), so single-seed comparison would be statistically fragile. Validation rank also fails to predict test rank—the best-validation seed 42 is not the best-test seed 123—consistent with the ≈10-pt val–test gap induced by geographically disjoint splits, a gap invariant to loss, model size and seed. Pooling the top-5 checkpoints of all three seeds (15 checkpoints) with TTA-8 gives 76.18% macro-$F1$ with no dependence on a fortunate seed; TTA alone contributes

+0.15 pts within the ensemble, its largest per-class effect being +4.5 pts on Jellyfish. A scripted 10-seed protocol that trains, bundles and evaluates each configuration without manual intervention is in place to convert the headline number into a distribution.

Table 5. Seed variance (LDAM $\mathcal{C}$ = 0.5, TTA-8). All metrics in %.

Seed	Val □$_1$	Test □$_1$	mIoU	OA
0	86.60	74.81	64.98	88.92
42	86.76	75.93	65.72	87.84
123	86.31	**76.43**	**66.55**	88.69
mean ± std		75.72±0.68	—	—

3. Computational Efficiency

Table 6 places the accuracy results in the context of what they cost. TRITON reaches state-of-the-art macro-F_1 on a single Titan-class GPU rather than two workstation cards, with 3.35M parameters and a 335 MB peak VRAM footprint at inference. A single model sustains 184,033 px/s (22.1 ms per 4,096-pixel batch), so one 240 × 240 MADOS patch (57,600 pixels) is classified in ≈0.3 s. The deployed configuration performs 15 × 8 = 120 passes per pixel and drops to 1,534 px/s; as shown above, that 120× cost buys independence from seed choice rather than accuracy, so the single-seed model is preferable whenever latency matters.

The cost TRITON does not eliminate is displaced to the CPU. The 45 spatial descriptors take 809 ms per patch on a single core (37.8 min for the 2,803 MADOS patches) and increase the stored dataset size from 127 to 278 MB. The step is embarrassingly parallel and computed once per scene—under 5 min on 8 cores—but it is a genuine precondition, and unlike MariNeXt's end-to-end pipeline, it must be reimplemented and revalidated before transfer to another sensor. The trade is deployment-dependent:

Table 6. Computational cost of TRITON (measured on our hardware) against MariNeXt (as reported by its authors [1] and in the accompanying deployment report [17]). "n/r" = not reported.

Aspect	TRITON (ours)	MariNeXt

Paradigm	per-pixel (68 feat. + Transf.)	dense seg. (SegNeXt + VSCP)
GPU	1× TITAN RTX, 24 GB	2× RTX A5000, 24 GB ea.
Parameters	**3.35 M**	n/r
Batch size	4,096 px	n/r
Throughput (1 model)	0.18 M px/s	n/r
Throughput (deployed)	1.5 k px/s	n/r
Inference latency	22.1 ms / batch	n/r
Peak GPU memory	**335 MB**	n/r (≤ 24 GB)
Preprocessing	45 spatial feat., 809 ms/patch	VSCP + geometric TTA
Feature storage	+151 MB (127 → 278 MB)	n/a (end-to-end)
Macro-F_1	**76.18** (ens.) / 76.43 (best)	76.0
mIoU	**66.38**	64.3
Overall accuracy	88.73	**89.1**

TRITON is preferable where GPU count or memory binds, a dense end-to-end model where preprocessing must stay inside a single GPU graph.

VII. DISCUSSION AND CONCLUSIONS

Representation dominating the outcome: Comparing the classification metrics between TRITON (76.43%),) and the spectral-only Transformer (60.87%), there is a difference of ≈15.6 points in favor of TRITON. From this difference, ≈11.5 points are attributable to the pixel2vects features, and the remaining 3.6 points corresponds to seed selection and ensemble. This improves the RF and RF++ observations of the MADOS results [1]: explicitly, computed spatial descriptors carry most of the contextual information this classification task needs, in a form that keeps every model input physically interpretable. Margins, not gradients, repaired the long tail: Focal loss, despite per-class weighting and difficulty modulation, left the four rarest classes between 1.5% and 49.9% $F1$, where LDAM's margin mechanism recovered 38–47 points with no architectural change. Focal loss and LDAM are often treated as interchangeable imbalance remedies; on this benchmark they are not. Capacity, data and limitation: The 9.26M-parameter variant improved validation but not the test performance, and the ≈10-point val–test gap was invariant to every intervention: with geographically disjoint splits and 103-pixel rare classes, generalization is constrained by data, not by model expressiveness. Three limitations follow from our design choices: per-pixel predictions carry no explicit spatial coherence, so post-processing (e.g., morphological filtering or a CRF) is a natural extension not explored here; the spatial descriptors use fixed scales and window sizes inherited from the literature, their joint optimization was not attempted, and their extraction imposes a CPU stage (Section 3) that a dense end-to-end model does not need; and Natural Organic Material remains effectively unsolved, with no path to improvement short of additional labeled data.

By analyzing our results, our proposed pixel2vecs representation (68 features) that makes spectral, multiscale, textural and local-context evidence explicit, allows TRITON to match and even slightly exceed (76.43% vs. 76.0% macro-*F*1 and 66.55 vs. 64.3% mIoU for our best single model) a more complex state-of-the-art segmentation pipelines in the literature. Our system runs on one graphic card using only 335MB (out of 24GB) of memory for the inference footprint, instead of two as in [1]. LDAM margin loss proved to be the only imbalance mechanism worth adding to our architecture. What still remains a challenge, is the dataset imbalance with some classes represented by few classified pixels, rather the model itself. Natural Organic Material (31 test pixels) and the dark-surface confusions among Oil Spill, Waves & Wakes and Jellyfish persisted across every configuration we tried, and the ≈10-point validation–test gap induced by geographically disjoint splits was invariant to loss, capacity and seed. For operational monitoring the outcome is a system whose errors are physically interpretable.

REFERENCES

[1]. K. Kikaki, I. Kakogeorgiou, I. Hoteit, and K. Karantzalos, "Detecting marine pollutants and sea surface features with deep learning in Sentinel-2 imagery," ISPRS Journal of Photogrammetry and Remote Sensing, 2024.

[2]. E. Dalsasso, M. Rußwurm, C. Donner, S. Darmon, R. de Vries, M. Volpi, and D. Tuia, "Self-supervised pre-training enables marine debris detection across sensors," Remote Sensing of Environment, vol. 339, p. 115391, 2026.

[3]. S. Bouchelaghem, L. Mamen, M. Balsi, A. Tibermacine, M. Moroni, and I. E. Tibermacine, "Binary reformulation for marine debris detection in Sentinel-2 imagery: An empirical study on extreme class imbalance using the first benchmarks on combined MARIDA and MADOS datasets," Frontiers in Marine Science, vol. 13, p. 1765021, 2026.

[4]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5]. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[6]. M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in Advances in Neural Information Processing Systems (NeurIPS), 2022.

[7]. G. Giannopoulos, M. Kremezi, V. Karathanassi, V. Andronis, D. Bliziotis, K. Kikaki, A. S. Oliveira, and A. Müting, "Two-stage oil spill detection in SAR using a domain-adapted Segment Anything model," Remote Sensing, vol. 18, no. 12, p. 1948, 2026.

[8]. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[9]. K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in Advances in Neural Information Processing Systems (NeurIPS), 2019.

[10]. A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in International Conference on Learning Representations (ICLR), 2021.

[11]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[12]. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models

for tabular data," in Advances in Neural Information Processing Systems (NeurIPS), 2021.

[13]. X. Wu, T. Arshad, and B. Peng, "Spectral spatial window attention transformer for hyperspectral image classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1–13, 2025.

[14]. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610–621, 1973.

[15]. T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[16]. T. Lindeberg, "Feature detection with automatic scale selection," International Journal of Computer Vision, vol. 30, no. 2, pp. 79–116, 1998.

[17]. NVIDIA Developer Blog, "High-tech AI framework transforms global marine pollution tracking," https://developer.nvidia.com/blog/high-tech-ai-fra mework-transforms-global-marine-pollution-tracking/, 2024, accessed: 2026-08-09.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Research Program: Computer-Aided Molecular Target Identification for Tuftsin-Phosphorylcholine in Autoimmune Diseases

Oleg Kupervasser
Department of Mathematics
Shamoon College of Engineering
Be'er-Sheva, Israel
olegku@sce.ac.il

Marina Bershadsky
Department of Computer Science
Shamoon College of Engineering
Be'er-Sheva, Israel
marinab@sce.ac.il

Gilad Halpert
Zabludowicz Center for Autoimmune Diseases
Sheba Medical Center
Ramat-Gan, Israel
gilad.halpert@sheba.health.gov.il

Yehuda Shoenfeld
Zabludowicz Center for Autoimmune Diseases
Sheba Medical Center
Ramat-Gan, Israel
yehuda.shoenfeld@sheba.health.gov.il

Vadim Povolotskii
Department of Software Engineering
Shamoon College of Engineering
Be'er-Sheva, Israel
vpovolotskii@gmail.com

Mariia Lishnevskii
Department of Software Engineering
Shamoon College of Engineering
Be'er-Sheva, Israel
mariako2@ac.sce.ac.il

Abstract
Tuftsin-Phosphorylcholine (TPC) is a potent immunomodulatory ligand known to suppress pathological immune responses, yet its precise molecular targets at the cell membrane remain partially characterized. To accelerate target discovery, we present a dual silico framework that integrates physics-based simulation with statistical learning. The first method uses Molecular Dynamics (MD) and flexible docking to compute explicit binding affinities between TPC and candidate active centers, based on empirical force fields and classical physical interactions. The second method employs a reverse Quantitative Structure-Activity Relationship (QSAR) approach, combined with machine learning, to identify the essential topological and electronic features of active centers that are characteristic of strong TPC binding. This dual computational pipeline establishes a predictive platform to rapidly map the precise mechanism by which TPC prevents immune attacks, laying the groundwork for targeted, steroid-free alternative therapies for managing autoimmunity.

Keywords: computer-aided target identification, tuftsin-phosphorylcholine, reverse QSAR, molecular dynamics simulations, empirical potential energy, classical force fields, ligand-receptor docking, machine learning algorithms, active center topology, allosteric inhibition cavities, electrostatic pockets, immunomodulatory ligands, autoimmune attack suppression.

I. SCIENTIFIC BACKGROUND AND STATE OF THE ART

Globular proteins catalyze viral or bacterial multiplication. Such a protein has the same site called an active center. Proteins function using this active center. The active center is the "hands" of the protein. If we want to inhibit the multiplication of a virus or a bacterium, we need to find a small molecule (called a ligand) that can strongly bind to the active center and thus stop (inhibit) the function of the protein. Such a small molecule will be a drug against viruses or bacteria. Chemists usually search for such molecules by synthesis and by investigating their inhibitory function in vitro. Obviously, such a wet-lab job is slow and expensive. However,

there is a way to do it much faster and more cheaply. It is creating drugs using a computer. Indeed, if we know the structure of the protein and the location of its active center, we can calculate (based on known physical laws of molecular interaction) the strength of attraction between the ligand and the active center of the protein using a computer (Figure 1 in [1-4]).

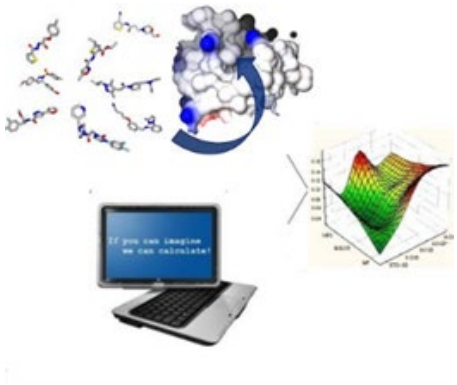


Figure 1. Computer-aided drug design.

In [4], we can find this first method by using Molecular Dynamics (MD) with Classical Potentials: "The central idea behind MD simulations is to study the time-dependent behavior of microscopic systems. This is obtained by solving the second-order differential equations represented by Newton's second law:

$$f_i(t) = m_i a_i(t) = -\frac{\partial V(x(t))}{\partial x_i(t)} \tag{1}$$

.re $f_i(t)$ is the net force acting on the iThe atom of the system at a given point in time $t, a_i(t)$ is the corresponding acceleration, and m_i is the mass. In Eq. (1), the instantaneous configuration of the system is represented by the vector. $x(t)$, which describes the position of the N interacting atoms in Cartesian space.

$$(x = \{x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_N, y_N, z_N\}) \tag{2}$$

Usually, in computational drug discovery, we adopt a classical mechanics description of the forces. Notably, this approximation holds for massive particles such as nuclei, whereas electron motion must be averaged out. To achieve this, an empirical potential energy function is introduced ($V(x)$ in Eq. (1)), and the model arising from this simplified representation is referred to as the force field (FF), or molecular mechanics (MM):

$$V = A + B + C + D + E, \tag{3}$$

where each term describes a specific physical component of the atomic interactions:

- Stretching of chemical bonds:

$$A = \sum_i^{bonds} \quad \frac{k_{l,i}}{2}\left(l_i - l_{0,i}\right)^2 \tag{4}$$

- Bending of valence angles:

$$B = \sum_i^{angles} \quad \frac{k_{\alpha,i}}{2}\left(\alpha_i - \alpha_{0,i}\right)^2 \tag{5}$$

- Torsional rotation around bonds:

$$C = \sum_i^{torsions} \quad \left\{\sum_k^M \quad \frac{V_{ik}}{2}\left[1 + cos\left(n_{ik} \cdot \theta_{ik} - \theta_{0,ik}\right)\right]\right\} \tag{6}$$

- Van der Waals interactions (Lennard-Jones potential):

$$D = \sum_{i,j}^{pairs} \quad \varepsilon_{ij}\left[\left(\frac{r_{0,ij}}{r_{ij}}\right)^{12} - \left(\frac{r_{0,ij}}{r_{ij}}\right)^{6}\right] \tag{7}$$

- Electrostatic Coulomb forces:

$$E = \sum_{i,j}^{pairs} \quad \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \tag{8}$$

In Eq. (3), the first three terms represent intramolecular interactions of the atoms. They describe variations in potential energy as functions of bond stretching, bending, and torsion between atoms directly involved in bonding. Summations represent them over bond lengths (l), angles (α), and dihedral angles (θ), respectively. The stretching and bending contributions share the same functional form, as harmonic potentials describe them using reference values. l_0 and α_0 and force constants k_l and k_α, respectively. However, because of their intrinsic periodicity, torsional terms are naturally defined by a cosine series of M terms for each dihedral angle. Thus, n_{ik} is a parameter describing the multiplicity for the kThe term of the series, $\theta_{0,ik}$, i_k is the corresponding phase angle, and V_{ik} is the energy barrier. Taken as a whole, this group of contributions is usually referred to as the "bonded" terms of the FF."

The modern theory of differential equations in biology is described in [5-7].

In drug design, we have to reach stable equilibria; only in the case of stable equilibria is the process of creating a medicine physically achievable. This stable equilibrium is at the global energy minimum, as given by Eq. (3).

Mathematical methods for optimal search (global minimum) are described in the papers [8-11].

In [4]: "Although FFs can reduce the computational burden, Eq. (1) can still be analytically solved for only a few atoms. For biological macromolecule simulations, numerical methods must be used to split the integration of the equations of motion into discrete time intervals, called time-steps, δt. In doing so, forces are assumed to be constants during each integration step. However, as reported in Eq. (1), the forces depend on atomic positions that change over time. Thus, a small δtguarantees reliable forces over time. In practice, with appropriate workarounds, a time step of 1 or 2 fs can be safely considered as a good compromise between calculation accuracy and efficiency. This is commonly achieved, for example, with the SHAKE and related algorithms. Here, the equations of motion are solved while still satisfying the geometric constraint for the fastest degrees of freedom (i.e., the stretching and bending of bonds involving hydrogen atoms). Other strategies envision, for example, repartitioning the mass of heavy atoms onto the neighboring hydrogen atoms, allowing for a time-step increase of up to 4 As an example of an integrator, the velocity-Verlet is a simple and widely used algorithm in MD codes. Within such an integrator, the positions at time $(t + \delta t)$ are calculated by current positions, velocities ($v(t)$), and accelerations according to

$$x_i(t + \delta t) = x_i(t) + v_i(t)\delta t + \frac{1}{2}a_i(t)\delta t^2 \tag{9}$$

Accelerations are, in turn, calculated by the forces acting on each atom, that is, by taking the first derivative of the potential energy with respect to positions with opposite sign as shown in Eq. (1). Then, velocities are propagated as follows:

$$v_i(t + \delta t) = v_i(t) + \frac{1}{2}[a_i(t) + a_i(t + \delta t)]\delta t \tag{10}$$

As seen in Eq. (10), the second term on the right-hand side corresponds to the arithmetic mean of the accelerations taken at time t and $(t + \delta t)$. Moreover, to calculate accelerations at the next step, positions must be updated *before* advancing velocities. In fact, to be used safely in an MD simulation, any numerical integrator

must satisfy some practical and theoretical requirements. They must follow the analytical trajectory as closely as possible and preserve the physical properties of the equations of motion, with total energy as a constant of motion, given by the sum of the potential and kinetic energy. $\big(K(p)\big)$:

$$H(x,p) = V(x) + K(p) \tag{11}$$

Here, $H(,p)$ is the classical Hamiltonian of the system that depends on the coordinates and momenta of the particles and returns the total energy (E) of the system. From these considerations, it should be clear that as long as the integrator works properly, MD naturally follows the motion of a microscopic isolated system, where neither matter nor energy is exchanged with the surroundings."

The second method is the Quantitative Structure-Activity Relationship ($QSAR$). Indeed, if there are many known pairs of "active centers" - "ligands" - with strong attraction, then it is possible (using machine learning) to find which features of the pairs are characteristic of such strong attraction. (Figure 2, in [12-15])

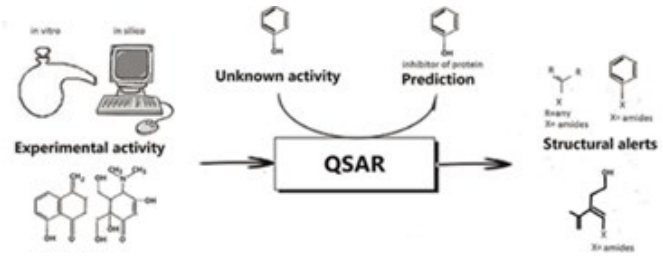


Figure 2. QSAR (*Quantitative Structure-Activity Relationship*)

In [14], we can find the history and development of $QSAR$: "The concept of a quantitative structure-activity relation. Until now, various $QSAR$Approaches have been developed gradually and have served as valuable predictive tools, particularly in the design of pharmaceuticals and agrochemicals. The first equation $QSAR$ was published by Crum-Brown and Fraser in 1868 (Brown et al. 1868), which is considered the first formulation of a quantitative structure activity relationship: the physiological activity" Φ was expressed as a function of the chemical structure C.

$$\Phi = f(C) \tag{12}$$

Studies were further extended by Richert et al. (1983), who reported that the toxicity of organic compounds is inversely related to their water solubility; in other words, differences in biological activities are due to changes in chemical and, especially, physicochemical properties.

$$\Delta\Phi = f(\Delta C) \tag{13}$$

The first well-defined $QSAR$ was developed and reported by Fujita and Ban (1970). Recently, various techniques have been introduced, and various software is available based on different algorithms in marketing."

II. RESEARCH OBJECTIVES AND SPECIFIC AIMS

We will leverage our expertise in computer programming, computer-aided molecular design, and QSAR to identify the specific active centers on the surfaces of immune cells that are blocked by TPC [16].

The primary objectives and the practical value of uncovering this mechanism are:

- Deciphering the Mode of Action: Uncovering exactly which active centers TPC blocks will explain *why* it successfully aborts immune attacks across multiple autoimmune models (lupus, arthritis, colitis, multiple sclerosis).
- Predicting Side Effects: Mapping all matching active centers on cell membranes will help identify off-target binding to non-intended proteins, ensuring drug safety.
- Rational Optimization: Revealing the precise geometry of the active center will enable chemists to modify the TPC molecule to make it more potent and selective (targeted design).
- Personalized Medicine: Disclosing specific target receptors will allow clinicians to match TPC therapy specifically to patients who overexpress these particular cell-surface proteins.

Our computational tools will establish an analytical platform to map how small immunomodulatory molecules suppress incorrect immune attacks and regulate human defense pathways.

III. DETAILED DESCRIPTION OF THE PROPOSED RESEARCH

1. Working hypothesis

We hypothesize that the interaction between the known drug TPC and its unknown targets on the surface of immune cells can be effectively modeled using physical and statistical algorithms. This will enable us to map the mechanisms of immune attack suppression significantly faster and cheaper than traditional in vitro proteomics. This predictive framework aligns with recent breakthroughs that combine neural networks and physical simulations in molecular structural biology, such as AlphaFold-3 [17] and Graph Neural Networks (GNNs) [18].

2. Research design and methods

Given the structure of TPC, we will calculate its binding affinity across thousands of documented active centers of human immune proteins using a computer. Reverse QSAR will use machine learning to identify structural and electronic features shared by proteins that strongly interact with TPC.

3. Analysis of potential active center locations on the cell membrane

Based on the foundational experimental data reported by Blank & Shoenfeld (2024), we identify the following critical zones for conducting a global energy minimum search:

- Allosteric inhibition cavities of innate immune receptors (e.g., TLR4): It is experimentally established that the phosphorylcholine (PC) component of TPC blocks the TLR4 signaling pathway, preventing the onset of an immune attack. Computational screening will focus on the inner hydrophobic cavity of the MD-2 co-receptor protein and allosteric side pockets at the TLR4/MD-2 interface. Binding of TPC to these zones is expected to induce a conformational shift that blocks the signal.
- Deep electrostatic pockets of navigational receptors (e.g., Neuropilin-1): Structural data indicate that the tuftsin domain of TPC binds to the Neuropilin-1 receptor, shifting the cell from an attacking state to a protective phenotype. Modeling includes calculating the insertion parameters of the TPC C-terminal arginine (R) into the deep, negatively charged cavity of the receptor, forming a hydrogen bond with the D320 residue of Neuropilin-1.

- Allosteric and boundary zones of presentation platforms (MHC / HLA Class I and II): Since TPC is too small to fit the classical peptide-binding groove of MHC, the software will scan the terminal barrier pockets of the MHC II groove (P1 and P9 pockets) for electrostatic shielding properties. Additionally, the outer loops of the MHC-TCR interface will be assessed, where TPC might act as a "molecular shield" to arrest the immune attack driven by T-lymphocytes.

4. Preliminary results and team competencies

- PI Dr. Oleg Kupervasser (Israel) possesses extensive expertise in chemical physics (Ph.D. from the Weizmann Institute) and postdoctoral experience in QSAR and 3D protein structures (University of Haifa). He developed software to locate loops and locks in protein configurations and to calculate the effects of surrounding solutions on protein-ligand interactions (the patented DISOLV program). This background is perfectly aligned with modeling how TPC fits into the active centers of immune receptors.
- Co-PI — Prof. Yehuda Shoenfeld (Israel) is a world-renowned authority on autoimmune diseases and the founder of the Zabludowicz Center for Autoimmune Diseases (the largest center of its kind in Israel). His unparalleled expertise in immune system pathophysiology ensures the project's fundamental biological relevance.
- Co-PI — Dr. Marina Bershadsky (Israel) is a leading expert in mathematics and software engineering, specializing in the modeling of complex biological systems and differential equations. Her contribution is vital for the real-time simulation of dynamic receptor behaviors.
- Vadim Povolotskii and Mariia Lishnevskii are Computer Software Engineering students at the Shamoon College.

In the laboratory, we utilize modern computational infrastructure.

IV. SIGNIFICANCE, IMPORTANCE, INNOVATION, AND POTENTIAL BENEFITS OF THE PROPOSED RESEARCH

The current status quo is the search for new drugs in vitro, drawing on prior experience without clear algorithms. We want to develop a universal, clear algorithm written in silico (i.e., in computer programs).

The current status quo in pharmacology relies on tracking drug mechanisms through prolonged *in vitro* screening. We propose a universal, clear, in silico algorithm to discover active centers that are blocked by a drug. The fundamental scientific breakthrough of this project is as follows:

If the hypothesis is validated *in silico*, this project will reveal a completely new class of small-molecule inhibitors and allosteric modulators of MHC (HLA) class I and II proteins. This will challenge the established dogma in classical immunology, stating that large presentation platforms of adaptive immunity are inaccessible to precise small-molecule intervention. We will shift from empirical observation of TPC effects to precise physical mapping of its hidden biophysical targets, discovering new allosteric pockets on the surface of key human immune receptors.

V. APPLICABILITY

Expected results and long-term therapeutic impact: The deployment of this software platform will provide a cost-effective, rapid method for identifying active centers in immune cells that TPC blocks. Defining the exact topographic coordinates ("cavities") on the MHC surface where TPC binding conformationally shuts down autoantigen presentation will lay the groundwork for developing potential alternatives to conventional steroid therapies. Unlike non-specific corticosteroids (such as dexamethasone or methylprednisolone), which cause severe systemic side effects and globally suppress the entire immune system, selective small-molecule TPC-based inhibitors are hypothesized to target the initiation step of the pathological immune response, potentially minimizing systemic adverse effects.

This will enable the effective treatment of severe autoimmune conditions while preserving the patient's general infectious safety and overall health. Risks and Mitigation: The previously identified risk of being unable to validate *in silico* findings without direct access to a wet laboratory is now completely mitigated. Prof. Yehuda Shoenfeld and his research group will provide full laboratory validation of the computational modeling results.

VI. COOPERATION VECTORS BETWEEN THE RESEARCH GROUPS

- In Israel (Computational PI — Dr. Oleg Kupervasser): Responsible for leading the development of mathematical algorithms and software for target searching, leveraging extensive experience in QSAR and molecular mechanics.
- In Israel (Biological Co-PI — Prof. Yehuda Shoenfeld): Responsible for the scientific supervision of the medical and biological validation phases. Prof. Shoenfeld will lead laboratory testing of modeling results (in vitro binding assays and cellular assays on macrophages and T-cells) to experimentally confirm the highly specific binding of TPC to the computationally predicted active centers.
- Mathematical Group (Co-PI — Dr. M. Bershadsky (Israel), Vadim Povolotskii (Israel), Mariia Lishnevskii (Israel)): Responsible for delivering mathematical optimization for tracking TPC-receptor interactions via molecular dynamics and implementing global optimization algorithms to reduce computational overhead.

VII. WORK PLAN AND GANTT CHART

Table 1. 3-Year Parallel Joint Research Roadmap Gantt Chart

#	Task Description	Year 1	Year 2	Year 3
1	Months 1–12 (Year 1): Computational group formulates the mathematical core, combining physical integrators and statistical QSAR learning rules for machine learning. Simultaneously, Prof. Shoenfeld's biology laboratory performs parallel *in vitro* testing of the baseline TPC molecule to provide experimental constants for method calibration.			
2	Months 13–24 (Year 2): The computational group optimizes the software of the universal method based on the combination of physical force fields and neural network recognition (inspired by the Nobel Prize in Chemistry 2024) and designs new candidates			

#	Task Description	Year 1	Year 2	Year 3
	ligands. The biology laboratory maintains parallel cell cultures and prepares validation platforms.			
3	Months 25–36 (Year 3): Computational group performs final scoring of the universal method. Biology laboratory executes parallel *in vitro* comparative assays (TPC vs new candidate ligands) to experimentally confirm the accuracy and predictive power of the developed universal method.			

REFERENCES

[1]. [1] Application of the docking program Sol for CSAR benchmark/A. V. Sulimov, D. C. Kutov, I. V. Oferkin et al.// Journal of Chemical Information and Modeling, 2013, vol. 53, pp. 1946 –1956.

[2]. [2] PCM, COSMO, and Surface GB implementation in program DISOLV: Development and Validation (USTNYJ), Sulimov V.B., Ofjorkin I.V., Kochugaeva M.P., Katkova E.V., 244th ACS National Meeting, Philadelphia, Pennsylvania, USA, 2012

[3]. [3] Novel Gridless Program SOL-P for Flexible Ligand Docking with Moveable Protein Atoms, Sulimov A., Zheltkov D., Sulimov V.B., Kutov D., Tyrtyshnikov E., Oferkin I. 21st European Symposium on Quantitative Structure-Activity Relationship: Where Molecular Simulations Meet Drug Discovery, Verona, Italy, September 4-8, 2016

[4]. [4] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli, Role of Molecular Dynamics and Related Methods in Drug Discovery, J. Med. Chem. 2016, 59, 9, pp. 4035–4061, January 25, 2016

[5]. [5] M. Bershadsky, M. Chirkov, A. Domoshnitsky, S. Rusakov, and I. Volinsky, ”Distributed Control and the Lyapunov Characteristic Exponents in the Model of Infectious Diseases”, Complexity, Volume 2019, Article ID 5234854, 12 pages (Published: November 13, 2019)

[6]. [6] A. Domoshnitsky, I. Volinsky, M. Bershadsky, ”Around the Model of Infectious Disease: The Cauchy Matrix and Its Properties”, Symmetry 2019, 11(8),1016; (Published: August 6, 2019)

[7]. [7] M. Bershadsky, L. Shaikhet, ”Stability Analysis of a Mathematical Model for Infectious Diseases with Stochastic Perturbations,” Mathematics, 13(14), 2265 (2025). https://doi.org/10.3390/math13142265

[8]. [8] M. A. Abramson, J. E. Dennis Jr., O. Brezhneva, and R. Pingel, Pattern Search in the Presence of Degenerate Linear Constraints, *Optimization Methods and Software*, Vol. 23(3), 2008, pp. 297–319.

[9]. [9] O. Brezhneva and A. A. Tret’yakov, When the Karush–Kuhn–Tucker Theorem Fails: Constraint Qualifications and Higher-Order Optimality Conditions for Degenerate Optimization Problems, *Journal of Optimization Theory and Applications*, Vol. 174(2), 2017, pp. 367–387.

[10]. [10] O. Brezhneva, E. Bednarczuk, K. Leśniewski, A. Prusinska, and A. A. Tret’yakov, Towards Nonlinearity: The p-Regularity Theory, *Entropy*, Vol. 27, Art. 518, 2025, 46 pages.

[11]. [11] O. Brezhneva, Y. Evtushenko, V. Malkova, and A. A. Tret’yakov, A p-Factor Interpolation Polynomial for Trust-Region Frameworks Near Degenerate Solutions, *Lecture Notes in Computer Science*, Vol. 16426, 2026, pp. 3–17, Springer.

[12]. [12] Ramensky V, Sobol A, Zaitseva N, Rubinov A, Zosimov V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. Proteins: Structure, Function, and Bioinformatics. 2007; 69 (2): 349-357

[13]. [13] Nikitin S, Zaitseva N, Demina O, Solovieva V, Mazin E, Mikhalev S, et al. A very large diversity space of synthetically accessible compounds for use with drug design programs. Journal of Computer-Aided Molecular Design. 2005; 19:47 - 63

[14]. [14] Ankur Vaidya, Sourabh Jain, Shweta Jain, Abhishek K. Jain, and Ram K. Agrawal, Quantitative Structure-Activity Relationships: A Novel Approach of Drug Design and Discovery, Journal of Pharmaceutical Sciences and Pharmacology, Vol. 1, 219–232, 2014

[15]. [15] Shailesh G. Jawarkar, Madhuri D. Game, Quantitative structure activity relationship: A tool for new drug design, World Journal of Pharmaceutical Sciences, ISSN (Print): 2321-3310; ISSN (Online): 2321-3086, 01-08-2018 http://www.wjpsonline.org/

[16]. [16] Miri Blank, Yehuda Shoenfeld, Helminth derivative tuftsin-phosphorylcholine to treat autoimmunity, Autoimmunity Reviews 23 (2024) 103601.

[17]. [17] Abramson J., et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, Vol. 630, 2024, pp. 493–500.

[18]. [18] Odin Zhang, et al. Graph Neural Networks in Modern AI-aided Drug Discovery. arXiv:2506.06915 [q-bio.BM] (2025)

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Design of a Compact Asymmetric Biconical Antenna for UAV Applications

Ali Larbah
Electrical and Electronic Engineering Department
Engineering College, Misurata University
Misurata, Libya
larbah.ali@eng.misuratau.edu.ly
ORCID: 0000-0002-2254-8363

Jalal Srar
Electrical and Electronic Engineering Department
Engineering College, Misurata University
Misurata, Libya
jalal.srar@eng.misuratau.edu.ly
ORCID: 0000-0002-9109-5556

Hassan Aldeeb
Electrical and Electronic Engineering Department
Engineering College, Misurata University
Misurata, Libya
haldeeb@eng.misuratau.edu.ly
ORCID: 0000-0001-9213-3852

Salem Alkhair
Electrical and Electronic Engineering Department
Engineering College, Misurata University
Misurata, Libya
Salemalkr6@gmail.com

Mohammed Alzwawi
Electrical and Electronic Engineering Department
Engineering College, Misurata University
Misurata, Libya
mohammedelzwawe@gmail.com

Abstract
This paper presents the design, fabrication, and experimental characterization of a lightweight asymmetric double-cone antenna optimized for wideband operation. The proposed antenna incorporates a miniaturized reflector diameter, optimized conductor thickness, and an adjustable feed gap to achieve significant weight reduction while preserving broadband performance. The structure was modeled and optimized using CST Studio Suite, followed by prototype fabrication and measurement using a vector network analyzer (VNA). Key performance metrics, including the reflection coefficient (S_{11}), voltage standing wave ratio (VSWR), and input impedance, were evaluated for various feed-gap configurations, with a 2 mm gap providing the best overall performance. The fabricated antenna demonstrates a measured operating bandwidth of approximately 4–14 GHz, with minimum reflection coefficients of −46.95 dB at 6.84 GHz and −35.57 dB at 9.45 GHz. Stable VSWR characteristics are maintained across the operating band. A comparison between simulated and measured results confirms the validity of the proposed design. Furthermore, benchmarking against a commercially available wideband antenna from a leading global manufacturer highlights the improved impedance-matching performance achieved by the proposed structure. The results verify that the optimized asymmetric configuration effectively reduces weight without compromising electromagnetic performance, making it suitable for wideband communication systems and uncrewed aerial vehicle (UAV) applications.

Keywords: Biconical antenna, asymmetric design, compact antenna, wideband, impedance matching, UAV applications.

I. INTRODUCTION

Biconical antennas, consisting of two conical conductors connected at a common apex, are widely employed due to their broad impedance bandwidth, omnidirectional radiation characteristics, and simple geometric structure. These features make them suitable for numerous applications traditionally served by half-wave dipole

antennas, including electromagnetic compatibility (EMC) testing, broadband communication systems, and uncrewed aerial vehicles (UAVs) [1–3].

Despite these advantages, conventional biconical antenna designs face several challenges. These include relatively large physical dimensions, limited control over the main beam direction in the E-plane, and significant size and weight constraints, which are particularly critical in UAV-based systems [4]. Extending the operational range of UAV communication systems typically requires increasing the transmitted power at both the UAV and the ground station. This can be achieved using directional antennas, antenna arrays, or tracking antenna systems that focus radiation toward the UAV. However, such approaches are often complex and costly, introducing additional challenges. For example, highly directional antennas necessitate careful mitigation of both co-channel and cross-channel interference. In contrast, antenna arrays significantly increase system complexity, which is undesirable for lightweight and compact UAV platforms [4]. To address the size and weight limitations of biconical antennas in UAV applications, the authors in [2] proposed a miniaturized biconical antenna loaded with metallic disks and cylindrical elements. This configuration expands the impedance bandwidth from 200 to 965 MHz while maintaining stable radiation patterns and gain,achieving a reduction in antenna height of 288.4 mm and diameter of 175mm. In [3], an irregular biconical antenna composed of two dissimilar cones connected by an annular ring was introduced to achieve a nearly constant beam direction over the operating bandwidth. The E-plane beam direction can be controlled by adjusting the annular ring dimensions and the slope ratio between the cones. However, once installed on a UAV, beam reconfiguration becomes impractical, as it relies on fixed geometric parameters. Furthermore, the limited bandwidth reported in [3] was later extended in [5] through an asymmetric miniaturized biconical antenna operating over the 3.1–10.6 GHz frequency range. Similarly, a dual-band (2.45 GHz and 5.80 GHz. A circularly polarized biconical antenna incorporating parasitic parallelepiped elements was proposed in [6] to enhance bandwidth and overall performance. Additional modifications to the classical biconical structure, including the integration of rings and cylindrical elements, were presented in [7], resulting in an ultra-wideband operation spanning from 1.5 GHz to 41 GHz. Integrating the antenna into the drone airframe provides an effective solution to structural and spatial constraints while reducing overall system weight. Since payload capacity is a critical design parameter in uncrewed aerial vehicles (UAVs), minimizing antenna weight without degrading electromagnetic performance is essential. In this work, a lightweight asymmetric bicone antenna is proposed and optimized using CST Studio Suite, followed by fabrication and experimental validation. The proposed design achieves approximately a 33% reduction in weight compared to a conventional bicone antenna, while maintaining comparable impedance bandwidth and radiation characteristics to those of a commercially available counterpart from a leading global manufacturer.

II. ANTENNA DESIGN

The mathematical model of a biconical antenna provides a comprehensive framework for describing its electromagnetic behavior through key parameters and governing equations. This model is fundamental for analyzing antenna performance and for optimizing the design to meet specific application requirements.

1. Biconical Antenna Geometry

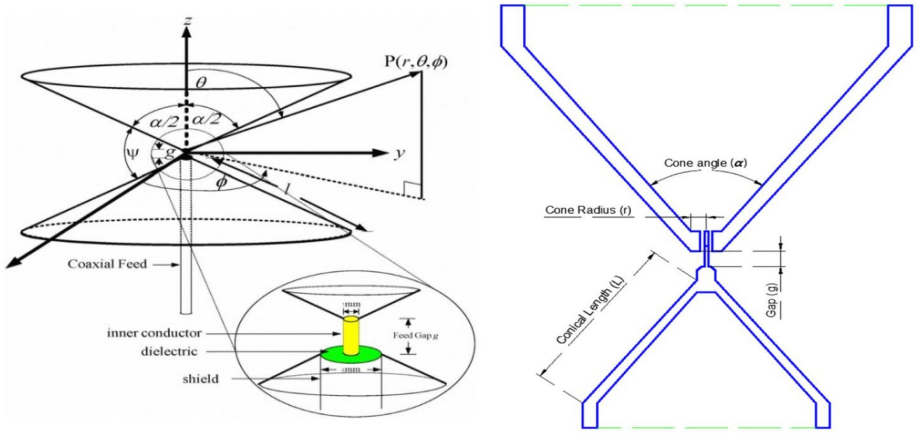
The proposed integrated and miniaturized conical dipole antenna employs an asymmetric configuration to enhance performance over a wide frequency range. The antenna consists of two conical sections that are positioned with a small gap between their apices. The first cone, located closer to the feed source or mounting structure, has a larger diameter, while the second cone, positioned farther from the source, has a smaller diameter. This asymmetric geometry is deliberately selected to improve impedance matching, bandwidth, and radiation characteristics.

Several geometric parameters govern the antenna's performance:

- The cone angle (α) primarily influences the antenna directivity and gain.
- The cone radius (r) affects the beamwidth and radiation resistance.
- The gap between the cones (g) plays a critical role in determining antenna gain and operational bandwidth.
- The cone length (L) controls the resonant frequency of the antenna.

As illustrated in Fig. 1, these parameters are treated as design variables and are systematically optimized to achieve the desired performance. The optimized antenna structure provides broadband operation with a realized gain ranging from approximately 2 to 6 dB. Furthermore, the return loss (S_{11}) is maintained below -10 dB across the operating frequency band, while the voltage standing wave ratio (VSWR) remains between 1 and 2 [8].

Figure 1(b) presents the final configuration of the proposed biconical antenna, which is derived from the classical structure shown in Figure 1(a). The modified design incorporates truncated cones, outer rings, a coaxial feed line, and cylindrical elements to enhance bandwidth and impedance matching, as illustrated in Figure 1(b) [5].



(a) (b)

Figure 1. The structure of the biconical antenna. (a) Geometry configuration, (b) Cross-section view

2. Input Impedance

For an infinite biconical antenna operating in free space, the characteristic impedance can be expressed as [9]

$$Z_o = 120ln\left(cot\left(\alpha/2\right)\right) \quad (1)$$

Since the characteristic impedance in (1) is independent of the radial distance r. It also represents the input impedance at the feed terminals of an infinite biconical structure. Moreover, because (1) corresponds to a purely resistive impedance, it can be further simplified for small cone angles. Under this assumption, the input impedance can be approximated as

$$Z_{in} \approx 120ln\left(4/\alpha\right) \quad (2)$$

During the optimization process, a 50-Ω coaxial feed line was connected to the antenna model. This configuration was adopted to account for the feeding mechanism accurately and to minimize the influence of cable effects on the antenna's impedance matching and overall performance.

3. Skin Depth Characteristic

Since the biconical antenna can be modeled as a transmission line, the thickness of the conductive material significantly affects its electrical performance. In this work, antenna weight is reduced by decreasing the thickness of the conductive sheets in the antenna structure. However, the minimum feasible thickness is limited by the skin depth. (δ) of the conductor.
For good conductors, the skin depth δcan be expressed as [10]

$$\delta = {1}/{\sqrt{\pi\delta\mu\sigma}} \tag{3}$$

where $f$as the operating frequency, $\sigma$is the electrical conductivity of the material, and $\mu = \mu_r\mu_0$is the magnetic permeability of the conductor. Here, $\mu_o = 4\pi \times 10^7\ H/m$ denotes the permeability of free space, and $\mu_r$is the relative permeability of the material.
It should be noted that (3) is valid only for good conductors, where conduction currents dominate over displacement currents. Table 1 summarizes the calculated skin depth values (δ) for the conductive materials employed in the fabricated antenna prototype.

Table 1. Maximum skin depth $\delta$of copper and aluminum biconical antennas at $f = 100$MHz.

	Copper	***Aluminum***
μ_r	0.9999912	1.00000037
σ	5.8×10^7	3.5×10^7
δ	7 μm	8.5 μm

Based on the skin depth (δ) values listed in Table 1, electromagnetic waves are effectively reflected from the antenna surface when the conductor thickness is significantly greater than δ. As reported in [11], within the operating frequency range of 1–13 GHz, The thickness of the conductive material has a negligible effect on the performance of a biconical antenna, provided that this condition is satisfied.

4. Size Re-Adjustment

Previous studies [12–14] have demonstrated a strong relationship between the physical dimensions of a biconical antenna and its radiation characteristics, particularly the E-plane radiation pattern. Systematic size-reduction targeting parameters such as cone length, flare angle, and material thickness has been widely adopted as an effective approach to antenna miniaturization and weight reduction. In this work, the antenna's weight was reduced by modifying its overall dimensions, including the reflector diameter and the conductor thickness. These adjustments inherently alter the critical dimensional ratios that define the antenna's asymmetric configuration. As noted in [12], changes to the upper section of the antenna—effectively altering the equivalent cap size and cone geometry—directly influence the downward tilt of the E-plane radiation pattern. Consequently, any dimensional reduction intended to reduce weight must be accompanied by a careful re-optimization of the antenna geometry to preserve the desired beam direction and overall radiation performance. The optimized antenna configuration was obtained using CST Studio Suite. Figure 2 shows the fabricated prototype based on this design. Its electromagnetic performance was subsequently validated through a series of experimental measurements, the results of which are presented and analyzed in this section III.



Figure 2. Fabricated an asymmetric biconical antenna after mechanical assembly.

III. RESULTS AND DISCUSSIONS

The proposed asymmetric biconical antenna was analyzed via simulations in CST Studio Suite and experimentally characterized with a Vector Network Analyzer (VNA). Key performance parameters, including the reflection coefficient (S_{11}Voltage standing wave ratio (VSWR), input impedance, and Smith chart were extracted to verify impedance matching across the targeted frequency band. In addition, the influence of varying the feed gap between the two conical elements on antenna performance was investigated. Simulated and measured results of the modified antenna were compared with those of the original copper and aluminum prototypes to assess performance variations. The antenna was characterized over the frequency range 0.8–18 GHz, including the targeted operating frequency of 1.2 GHz. For benchmarking purposes, identical measurement procedures were conducted on a commercially available dual-cone antenna from a global manufacturer. Furthermore, two gap lengths, 2 mm and 8 mm, were investigated to evaluate their influence on the antenna's performance. The feed gap (Gap) The distance between the two conical sections is a critical parameter affecting broadband performance. In this study, the performance of the manufactured antenna was evaluated for two gap lengths. The impact of the gap on impedance matching, resonant frequency, and operational bandwidth was assessed by analyzing S11 and SWR curves. Due to manufacturing constraints for the 2 mm gap, an iron adjustment piece was used to achieve the required spacing, whereas a copper piece was used for the 8 mm gap. It should be noted that these material selections were dictated solely by mechanical considerations for precise gap implementation, rather than by a comparison of metallic properties. Aluminum was unavailable at this stage for these adjustments.

Figure 3 illustrates the geometric configuration of the antenna for a 2 $mm\ gap, showing\ the\ position\ of\ the\ iron\ adjustment\ piece\ relative\ to\ the\ feed, while\ the\ 8\ mmThe$ gap configuration employs a copper piece. This visualization highlights the effect of short feed distances on impedance matching near the resonant frequency. This analysis aims to identify the optimal gap that provides the best combination of impedance matching and frequency stability for the fabricated antenna.

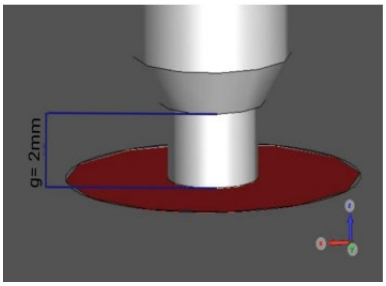
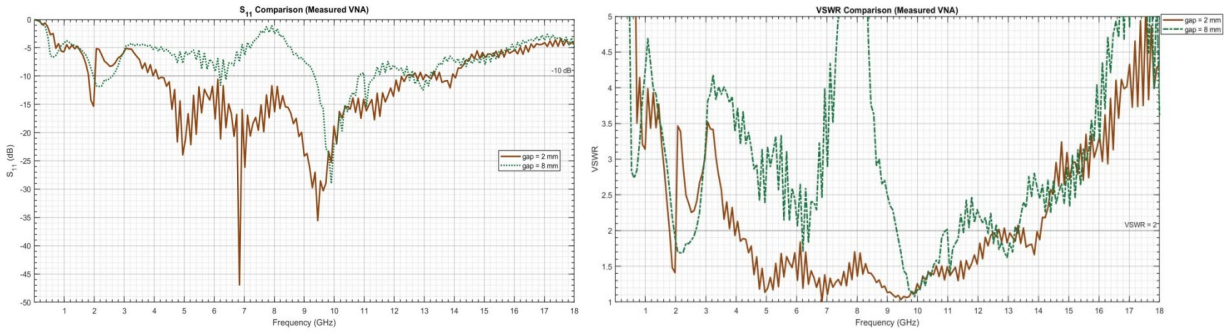


Figure 3. Feed gap of 2 mm implemented using an iron adjustment piece.

After conducting experimental measurements on the fabricated antenna with both gap lengths, the antenna performance was evaluated based on the measured reflection coefficient (S_{11}), bandwidth, and VSWR. Figure 4 presents the measured. $S_{11}$and VSWR curves over the frequency range of 0.8–18 GHz for the 2 mm and 8 mm gap configurations. The key observations are summarized below.

- From Figure 4(a), for the 2 mm feed gap, the lowest S_{11}values occur at 6.84 GHz and 9.45 GHz, with corresponding magnitudes of −46.95 dB and −35.57 dB, respectively. In contrast, for the 8 mm gap, S_{11}exhibits multiple fluctuations around the −10 dB level, with the minimum value of −28.85 dB observed at 9.9 GHz. This indicates that the 2 $mmThe\ The\ The\ equation\ gapThe$ gap provides consistently lower reflection coefficients across most of the measured frequency range.
- Applying the $S_{11} \leq -10\ dB$ criterion, the operational bandwidth for the 2 mm gap is approximately 4–14 GHz, resulting in a bandwidth of 10 GHz. Conversely, for the 8 mm gap, the fluctuating S_{11}Response makes it difficult to define a clear, continuous operational band that meets this matching criterion.
- Figure 4(b) presents the VSWR curves, which closely correspond to the behavior of S_{11}. For the 2 mm gap, ideal VSWR values near 1 are observed at 6.84 GHz and 9.45 GHz, indicating improved power transfer and reduced reflection at these frequencies. For the 8 mm gap, impedance matching is primarily effective near the resonant frequency of 9.9 GHz, while VSWR values increase outside this region, indicating higher reflected power across the remaining frequency band. Overall, the 2 mm gap exhibits a more stable and significantly lower VSWR over a wide frequency range compared to the 8 mm gap, confirming its superior impedance matching performance.



(a) S_{11} Measurements (b) VSWR Measurements

Figure 4. Measurements of $S_{11}$and VSWR for the fabricated antenna with 2 and 8 mm feed gaps.

To further evaluate impedance matching, the antenna's input impedance was plotted on a Smith chart using a Vector Network Analyzer. It should be noted that the trajectory on the measured Smith chart may appear unclear due to the large number of frequency points and minor ripples caused by measurement noise. Consequently, the optimal matching points were identified programmatically using MATLAB and treated as reference values.

Figure 5 shows the Smith chart for the fabricated antenna with m , e,p. Based on this analysis, the two best-matching points were found at 6.84 GHz and 9.45 $GHz$with corresponding input iimpedances of cap Z sub in equals 50.4 minus j 1.62 cap omega and cap Z sub in equals 49.56 minus j 0.09 cap omega respectively The proximity of these values to 50 Ω, combined with minimal imaginary components, confirms excellent impedance matching, consistent with tthe 11th cap S sub 11thVSWR measurements.

For the 8 mm gap, the Smith chart was not presented due to increased ripple and a less clear trajectory in the measurement data. Instead, the numerical values of Z_{in} The reference frequencies were used to assess matching performance.

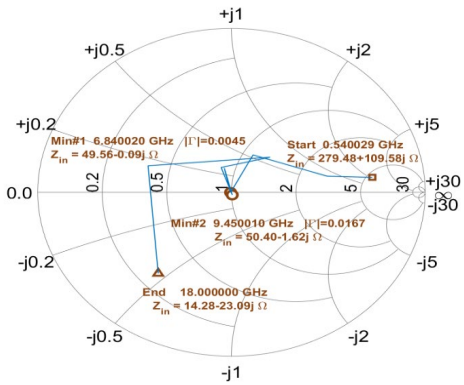


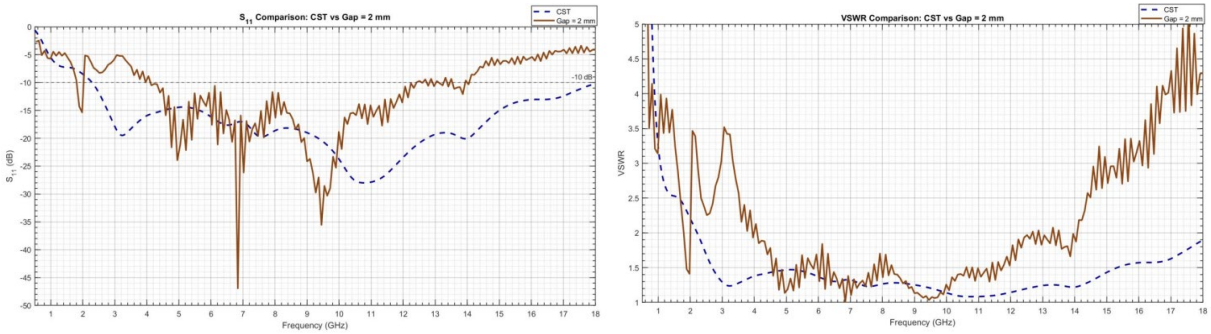
Figure 5. Smith chart of the fabricated antenna with a 2 mm feed gap.

1. Acronyms: Comparison Between Simulation and Practical Measurements

Based on the optimal practical tuning of the fabricated antenna with a 2 mm feed gap, the simulated results obtained using CST Studio Suite were compared with the antenna's measured performance. This comparison was performed using the reflection coefficient (S_{11}), VSWR curves, and the input impedance (Z_{in}) at specific frequencies. Figure 6(a) presents both the simulated and measured. S_{11}curves on a single plot to highlight the differences between the CST predictions and the practical measurements. The CST simulation curve exhibits a smooth response, whereas the measured data show distinct ripples due to fabrication tolerances, connector effects, and measurement noise. Additionally, slight shifts are observed in the positions of the main reflection dips. The practical measurements show deeper, multiple dips compared to the simulation, reflecting subtle real-world variations not captured by the idealized CST model. Similarly, the VSWR curves from the CST simulation and the practical measurements were compared on a single plot, as shown in Figure 6(b). This comparison corroborates the reflection behavior observed in the S_{11}Results and highlights the differences in fluctuation between the practical measurements and the simulation. The simulated curve exhibits a smoother response, while

the measured VSWR shows small ripples and variations due to fabrication tolerances, connector effects, and measurement noise, consistent with the behavior observed in the S_{11}Comparison.

Regarding the operational bandwidth, the CST simulation predicts a wider frequency span of approximately 2.3–18 GHz, whereas the practical measurements indicate an effective bandwidth of roughly 4–14 GHz. This discrepancy is primarily attributed to fabrication tolerances, connector effects, and minor variations in material properties, which slightly limit the measured performance compared to the idealized simulation model.



(a) (b)

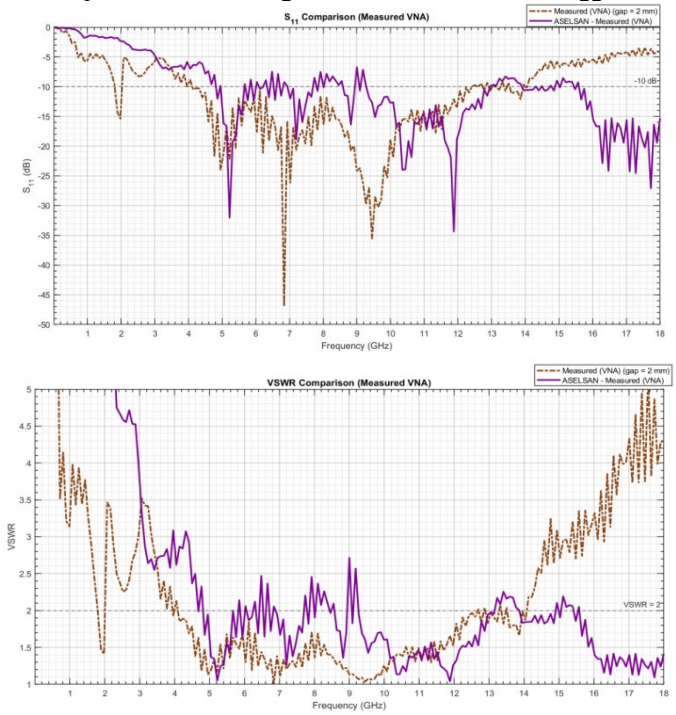
Figure 6. (a) Measured vs. CST simulation results for Gap = 2 mm,
VSWR comparison between CST simulation and practical measurement for Gap = 2 mm.

2. Comparison of Practical Measurements Between the Fabricated Antenna and a Commercial Benchmark Antenna

Experimental measurements were conducted on a commercially available dual-cone antenna from a global manufacturer using a vector network analyzer (VNA) to enable direct comparison with the locally fabricated antenna. The comparative analysis focuses on key performance parameters, including the reflection coefficient (S_{11}), operating bandwidth, and frequency response, to quantify the differences between the two designs.

- Figure 8 presents the measured reflection coefficient (S_{11}) for both antennas. The fabricated antenna exhibits deep reflection dips, with the most prominent occurring at 6.84 GHz (−46.95 dB) and another strong dip at 9.45 GHz (−35.57 dB). In comparison, the benchmark antenna shows a minimum reflection at different frequencies, primarily in the mid-frequency range, while the target operating frequency of 10.5 GHz falls within a region where both antennas achieve acceptable performance.
- Moreover, Figure 7(a) highlights that the fabricated antenna achieves an almost continuous operational bandwidth from approximately 4 to 14 GHz. By contrast, the reference antenna demonstrates good performance over a narrower mid-range frequency band, with additional improvements observed at higher frequencies. This comparison underscores the effectiveness of the proposed design in achieving a wide operational bandwidth while maintaining a minimum depth of reflection at key frequencies.
- Figure 7(b) presents the VSWR measurements for both the fabricated and commercial antennas. As expected, the reductions in VSWR correspond closely to the reflection minima observed in the S_{11}curves. The results indicate that VSWR values are lower and more stable across the mid-frequency range for both antennas. However, for the fabricated antenna, VSWR increases at higher frequencies,

reflecting a slight rise in reflected power outside the primary operational band. Overall, the VSWR behavior confirms the impedance matching trends observed in the S_{11}measurements.



(a) (b)

Figure 7. (a) Measured S_{11}, (b) Measured VSWR for fabricated and benchmark antennas.

IV. CONCLUSION

In this study, a lightweight asymmetric biconical antenna was designed, fabricated, and experimentally characterized for wideband applications. By optimizing key structural parameters, including reflector diameter, conductor thickness, and feed gap, the antenna weight was significantly reduced while maintaining broadband performance. The fabricated prototype with a 2 mm feed gap demonstrated excellent impedance matching, characterized by deep reflection minima at 6.84 GHz and 9.45 GHz, a wide measured operating bandwidth of approximately 4–14 GHz, and stable VSWR across the operating band. Good agreement between CST Studio Suite simulations and experimental measurements validates the proposed design methodology. Furthermore, benchmarking against a commercially available global antenna confirms the proposed structure's competitive performance in impedance matching and bandwidth. The results demonstrate that the proposed asymmetric configuration effectively balances weight reduction with high electromagnetic performance, making it a strong candidate for broadband communication systems and uncrewed aerial vehicle (UAV) applications. Future work will focus on further miniaturization, structural integration into UAV platforms, and the development of antenna array configurations to enhance directivity and gain.

ACKNOWLEDGEMENT

This research was conducted collaboratively within the Antennas and UAV Research Group at the Faculty of Engineering, Misurata University. The student authors jointly contributed to the antenna design, simulation, fabrication, experimental measurements, and manuscript preparation under the supervision of the faculty members.

REFERENCES

[1]. B. Amartya, P. Kaushik, C. Sayan, B. Gupta, and B. Anup, "Theoretical Investigations on CPW-Fed Single and Dual-Polarized Slot Radiators Using Schelkunoff's Biconical Antenna Analysis", Radio engineering, vol. 29, pp. 305-312, 2020, doi: 10.13164/re.2020.0305.

[2]. S. He, L. Chang, and Z. Chen, "Design of a Compact Biconical Antenna Loaded With Magnetic Dipoles," in IEEE Antennas and Wireless Propagation Letters, vol. 16, pp. 840-843, 2017, doi: 10.1109/LAWP.2016.2608920.

[3]. J. Ghalibafan, and S. Hashemi, "Design of an asymmetric capped biconical antenna for constant beam direction over a desired range of frequencies, AEU - International Journal of Electronics and Communications, vol. 84, pp. 27-33, 2018, https://doi.org/10.1016/j.aeue.2017.11.017.

[4]. A. Kelechi, M. Alsharif, D. Abdulbasit, P. Achimugu, O. Ubadike, J. Nebhen, A. Anthony, and P. Uthansakul, "The Recent Advancement in Unmanned Aerial Vehicle Tracking Antenna: A Review", Sensors (Basel), 2021, doi: 10.3390/s21165662.

[5]. Y. Zhiheng, R. Pengshan, M. Dengpan, and Z. Dong, "Design of a Compact Biconical Antenna with Asymmetric Configuration," 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 2018, pp. 1022-1026, doi: 10.1109/CompComm. 2018.8780684.

[6]. P. Janpangngern, D. Torrungrueng, M. Krairiksh, and C. Phongcharoenpanich, "Dual-Band Circularly Polarized Omni-Directional Biconical Antenna With Double-Circular Parallelepiped Elements for WLAN Applications," in IEEE Access, vol. 10, pp. 31970-31980, 2022, doi: 10.1109/ACCESS.2022.3160476.

[7]. S. Zhekov, A. Tatomirescu, and G. Pedersen, "Modified biconical antenna for ultrawideband applications," in 10th European Conference on Antennas and Propagation (EuCAP), Switzerland, pp. 1-5, 2016.

[8]. A. Bobreshov, A. Zhabin, E. Seregina, and G. Uskov, "Biconical antenna with inhomogeneous dielectric lens for UWB applications", Electronics Letters. Vol. 56, issue 17, pp. 857-859, Aug.2020.

[9]. Balanis, C. A. Antenna Theory: Analysis and Design. 3rd Edition. Wiley, 2016.

[10]. M. Sadiku, "Elements of Electromagnetics", 3rd edition, Oxford University Press, 2000.

[11]. J. Tuomela, B. Xue, K. Haneda, and C. Icheln, "Design and Fabrication Methods for a Biconical Antenna at 28 GHz ", 19th European Conference on Antennas and Propagation (EuCAP 2025), Stockholm, Sweden, April 2025.

[12]. Javad Ghalibafan, Seyed Mohammad Hashemi, "Design of an asymmetric capped biconical antenna for constant beam direction over a desired range of frequencies, AEU - International Journal of Electronics and Communications, vol. 84, pp.27-33, 2018.

[13]. R. Kudpik, K. Meksamoot, N. Siripon, and S. Kosulvit, "Design of a compact biconical antenna for UWB applications," 2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS), Chiang Mai, Thailand, 2011, pp. 1-6, doi: 10.1109/ISPACS.2011.6146212.

[14]. R. Janaswamy, "Input Admittance, Directivity and Quality Factor of Biconical Antenna of Arbitrary Cone Angle," IEEE Trans. Antennas Propag., vol. 70, no. 5, pp. 3248-3258, May 2022.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Leveraging AI Agents for Collaborative and Independent Authorship

Lamont Grissom
Dept. of Information Science
University of Arkansas at Little Rock
Little Rock, USA
lagrissom@ualr.edu

Daniel Berleant
Dept. of Information Science
University of Arkansas at Little Rock
Little Rock, USA
jdberleant@ualr.edu

Venkata Batthula
Department of Computer Science
Southwestern College
Winfield, KS, USA
Venkata.batthula@sckans.edu

Richard S. Segall
Dept. of Information Systems & Business Analytics
Arkansas State University Jonesboro, USA
rsegall@astate.edu

Abstract
Generative artificial intelligence systems are increasingly used to assist with long-form writing tasks, including educational, technical, and advisory content. However, single-agent large language models frequently struggle with maintaining narrative coherence, factual consistency, and structural integrity across extended documents, particularly in data-intensive domains. These limitations pose challenges for responsible and scalable AI-assisted authorship. This paper presents an exploratory, comparative case study of two AI-assisted authorship paradigms for producing a book-length manuscript on personal investing: (1) a human-in-the-loop, multi-persona collaborative authorship model implemented using Google's Gemini, and (2) an autonomous multi-agent system implemented using the LangGraph framework. Both approaches were assigned identical task constraints, outlines, and prompts to isolate differences attributable to the collaboration approach. The single-case comparison supported reporting the findings descriptively, supporting an initial understanding of the trade-offs of these two approaches. Results show that the human-in-the-loop Gemini model achieves superior narrative coherence, stylistic consistency, and advisory nuance, but requires substantial human coordination. The LangGraph multi-agent system demonstrates strong structural adherence and improved identification of factual claims but exhibits reduced narrative fluidity. This work contributes a comprehensive, reproducible methodology for evaluating multi-agent AI systems in long-form authorship, case study-based experiential evidence of trade-offs between autonomy and quality, and a hybrid agent architecture that can be extended and empirically validated by future researchers.

Keywords: Generative AI, Multi-Agent Systems, Human-in-the-Loop, Autonomous Authorship, Google Gemini, LangGraph, Long-Form Text Generation, Large Language Models (LLM).

I. INTRODUCTION

Since the public release of ChatGPT in late 2022, generative artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT, Gemini, and similar architectures, has dramatically reshaped how written content is produced, curated, and understood across domains ranging from creative writing to scientific authorship. These models have evolved from simple autocomplete tools into active collaborators capable of generating paragraphs, sections, and even complete documents with minimal human guidance. Research on the sociotechnical impacts of generative AI in authorship has noted that these systems alter traditional notions of text creation and creative agency, often resulting in hybrid human–machine outputs that challenge established academic practices and norms concerning authorship, responsibility, and disclosure [1], [2].

Despite their expressive power, LLMs continue to exhibit persistent limitations when tasked with generating sustained long-form content. While short-form outputs often demonstrate fluency and task relevance, extended documents pose challenges in maintaining global coherence, narrative continuity, and factual consistency, leading to topic drift and error propagation across sections [3], [4]. Hallucination -the confident generation of incorrect or ungrounded information - has been repeatedly documented as a significant reliability risk, particularly in structured and high-stakes domains such as finance, law, or medicine, where inaccurate claims can undermine trust and practical value [5], [6]. These shortcomings are exacerbated in single-agent systems that lack robust internal verification, context retention, or dynamic planning mechanisms, often resulting in incoherent narratives and unverifiable assertions over long spans of text [7], [8]. While architectural innovations such as retrieval-augmented generation (RAG) and enhanced planning have mitigated certain failure modes, they do not fully resolve issues of narrative cohesion or systematic consistency without extensive human feedback [9], [10]. To mitigate such issues, research has increasingly explored multi-agent AI architectures that distribute writing tasks across specialized subagents - such as planners, outline constructors, draft authors, editors, and critics - rather than relying on a single monolithic generative process [11], [12]. These frameworks aim to mirror human collaborative workflows and introduce internal checks that can improve reliability, traceability, and content quality beyond what is achievable with single-agent generation [13]. In parallel with architectural innovation, research on human–AI collaboration emphasizes structured interactions and feedback loops between human authors and generative systems to balance efficiency with oversight. Normative conversational frameworks and interactive oversight mechanisms have been proposed to align system outputs with human intent and disciplinary norms, improving task alignment and communication clarity [14], [15]. However, evaluations of multi-agent systems for complete, long-form nonfiction writing remain limited**.** Most existing studies either focus on short tasks, creative narratives in constrained settings, or modular benchmarks that do not capture the demands of sustained nonfiction discourse. There is thus a critical gap in the literature regarding (a) the scalability and coherence of multi-agent authorship pipelines, (b) their factual reliability in domains with rigorous accuracy requirements, and (c) comparative studies of human-in-the-loop versus fully autonomous multi-agent frameworks in realistic editorial workflows [4], [12], [6]. This paper addresses that gap with a case-study comparison of two paradigms of AI-assisted authorship: a human-in-the-loop, multi-persona collaborative model implemented with Gemini, and an autonomous multi-agent system implemented with LangGraph. The primary research question is whether multi-agent AI frameworks can generate coherent, accurate, and well-structured long-form written content with limited human intervention. Secondary questions examine the extent to which such systems replicate human collaborative writing processes and the trade-offs between autonomy, efficiency, and narrative quality [3], [11].

II. CONCEPTUAL AND THEORETICAL FOUNDATIONS

1. Knowledge Co-Creation and Human–AI Collaboration

This study is grounded in the theory of knowledge co-creation, which conceptualizes knowledge production as a continuous cycle between tacit and explicit knowledge. In generative AI writing environments, this cycle manifests as an iterative interaction among human intent, machine-generated text, and successive refinement. Distributed Cognition theory further frames cognition as distributed across individuals, artifacts, and environments. This principle is exemplified by tools that externalize otherwise implicit cognitive processes into inspectable, manipulable structures, such as visual representations of complex query logic [16]. Within AI-assisted authorship, LLMs and AI agents function as cognitive artifacts that share reasoning, memory, and organizational tasks with human collaborators, extending rather than replacing human cognition. Trust plays a central role in determining the effectiveness of such collaborations. In authorship contexts, trust governs when humans accept, revise, or override AI-generated text, shaping both workflow efficiency and final output quality.



2. Human-Centered AI and Socio-Technical Writing

Human-Centered AI emphasizes systems that amplify human abilities rather than automate them away. From this perspective, AI-assisted writing is best understood as a socio-technical process involving automation, human judgment, institutional norms, and ethical constraints. Writing itself is inherently socio-technical, involving tools, audiences, conventions, and accountability structures. When AI systems participate in writing, evaluation must therefore consider not only output quality but also process transparency, human labor, and ethical implications.

3. AI-Assisted Authorship and Multi-Agent Systems

Prior research has shown that single-agent LLM systems struggle with factual consistency and narrative continuity in extended documents, particularly in analytical or financial domains. Multi-agent debate and verification frameworks have been proposed to mitigate hallucinations by enabling internal critique among agents. Frameworks such as LangGraph operationalize these ideas by enabling structured agent workflows with defined roles and communication pathways, allowing AI systems to approximate human editorial collaboration.

III. STUDY OVERVIEW

This study employs a comparative case study design in which two AI-assisted authorship systems are tasked with producing a complete, book-length manuscript. A case study approach was selected because long-form writing involves interdependent decisions, iterative refinement, and emergent behaviors that are difficult to isolate using benchmark-style evaluations. Each case corresponds to a distinct authorship paradigm: (1) a human-in-the-loop, multi-persona collaborative system and (2) an autonomous multi-agent system. Both systems were evaluated under identical task constraints to ensure comparability.

IV. METHODOLOGY

1. Rationale for Case Study Design

This study adopts a single-case, qualitative comparative design. The design supported this exploratory investigation of how architectural differences between human-in-the-loop and autonomous multi-agent systems manifest in long-form, book-length writing. This design is suitable in part because long-form nonfiction authorship lacks a single correct output against which to evaluate system performance. That makes quantitative metrics such as accuracy scores against a reference text poorly defined for this task. Additionally, the theoretical constructs motivating this study, namely knowledge co-creation, distributed cognition, and trust, describe processes rather than discrete outcomes, and are more directly observable through a structured qualitative comparison of two authorship processes in a case study than through aggregate quantitative scoring. The comparative case study reported here is intended as an initial experiment to support future experiments that evaluate specific hypotheses quantitatively using a larger set of cases.

2. Task Definition

Both systems were tasked with generating a non-fiction manuscript of approximately 20,000 words on personal investing. The intended audience was defined as individuals aged 40 and older with limited retirement savings. This task was selected because it requires sustained narrative coherence, factual accuracy, ethical responsibility, and practical clarity across chapters, making it a stringent test of long-form AI authorship.

3. Systems Evaluated
4. Human-in-the-Loop Multi-Persona System (Gemini)

The human-in-the-loop system employed a single large language model operating under multiple conceptual personas instantiated through prompt conditioning. Personas represented roles such as researcher, writer, and editor. A human operator coordinated persona interactions, resolved contradictions, sequenced tasks, and integrated outputs. The human operator functioned as a centralized planner and editor, exercising judgment over content quality, narrative flow, and factual plausibility.

5. Autonomous Multi-Agent System (LangGraph)

The autonomous system was implemented using LangGraph, a Python-based agent orchestration framework. Agents operated as discrete computational entities with defined roles, message-passing rules, and execution order. After initial configuration, the system executed with minimal human oversight, emphasizing autonomy and internal coordination.

6. Agent Architecture

The autonomous system employed a hybrid hierarchical architecture that combined the Planner–Executor and Critic–Reviewer models. The Planner agent decomposed the manuscript outline into chapters and sections. Executor agents drafted content for assigned sections. A Critic agent evaluated factual consistency and logical structure, while an editor agent revised text for clarity, tone, and coherence. A Fact Extractor agent isolated explicit factual claims for review.

7. Workflow Coordination

Agent execution followed a directed workflow: planning, drafting, fact extraction, critique, editing, and reintegration. Termination conditions were defined as completion of all planned sections with no unresolved Critic flags. The system did not dynamically re-plan tasks unless manually reset, allowing controlled evaluation of workflow stability.

8. Prompting Strategy

Each agent was instantiated with a role-specific prompt defining responsibilities, evaluation criteria, output format, and constraints. Prompts were fixed across runs to minimize variance introduced by prompt drift.

9. Human Intervention Taxonomy

Human intervention was operationally defined and logged under five categories: structural intervention, factual correction, stylistic revision, error recovery, and prompt adjustment. Each intervention was time-stamped and categorized to enable comparison of human labor across systems.

10. Data Collection

Collected artifacts included raw agent outputs, final manuscript drafts, agent communication logs, prompt chains, intervention logs, time spent per phase, and token usage where available.

11. Evaluation Protocol

Evaluation focused on coherence, structural adherence, factual claim density, human intervention frequency, and process efficiency. Human reviewers assessed manuscripts using standardized rubrics, with AI-based scoring used as a secondary reference. We reported the roles, categories, and workflow structure of both systems above.

V. RESULTS

This section reports observations from the comparative evaluation of the human-in-the-loop Gemini system and the autonomous LangGraph multi-agent system. Results are organized by evaluation dimension and presented descriptively, summarized first in Table 1, followed by further details. Further interpretation is given in the following section.

Table 1. Comparative summary of observed findings across evaluation dimensions. Findings reflect a qualitative comparison of the two approaches used in our case study.

Evaluation Dimension	**Human-in-the-Loop Multi-Persona (Gemini)**	**Autonomous Multi-Agent (LangGraph)**
Narrative coherence & voice	High and consistent across chapters; effective long-horizon context management; smooth transitions with cross-referencing.	Locally coherent within sections but degrades globally, with templated transitions and declining voice consistency in later chapters.
Structural adherence	Occasional deliberate deviation from outline (merging/reordering), reflecting human judgment; strong within-section organization.	Near-perfect adherence to outline; minimal duplication/omission; uniform section lengths and formatting.
Factual claim density & accuracy	Fewer discrete claims, embedded in narrative; more contextualization/hedging; inaccuracies corrected more consistently by human oversight.	Higher density of explicit, enumerated claims via fact-extraction agent; critique phase flags inconsistencies, but flags are not always resolved.
Human intervention	Frequent but lightweight; proactive, occurring during drafting (sequencing, clarification, stylistic smoothing).	Infrequent but substantial; reactive, occurring after multiple agents pass (correcting accumulated repetition/disjunction).
Process efficiency & stability	Slower initial drafting; fewer large-scale rewrites; failures tied to human fatigue and coordination overhead.	Faster initial drafting via parallel execution, but longer revision cycles; failures include repetitive phrasing and critic-editor deadlocks.

1. Narrative Coherence and Voice Consistency

Manuscripts generated using the human-in-the-loop system exhibited high levels of narrative coherence across chapters. Reviewers observed consistent authorial voice, smooth transitions between sections, and frequent



cross-referencing of earlier concepts. Later chapters maintained continuity with earlier framing, indicating effective long-horizon context management.

In contrast, manuscripts produced by the autonomous multi-agent system were locally coherent at the section level but exhibited degradation in global narrative flow as document length increased. While individual sections were readable and logically structured, transitions between sections frequently relied on templated phrases and explicit structural markers rather than rhetorical continuity. Reviewer ratings for voice consistency declined across later chapters in the autonomous system, with repeated sentence constructions and stylistic motifs observed. This pattern was not observed in the human-in-the-loop condition.

2. Structural Adherence and Document Organization

The autonomous multi-agent system demonstrated near-perfect adherence to the predefined outline. All planned sections were generated in the intended order, with minimal duplication or omission. Section lengths were relatively uniform, reflecting planner-driven task decomposition. The human-in-the-loop system occasionally deviated from the original outline by merging or reordering subsections. These deviations were introduced deliberately by the human operator and typically reflected conceptual rather than structural considerations. Within-section organization was strong in both systems, though the autonomous system exhibited more consistent formatting and explicit sectional signaling.

3. Factual Claim Density and Accuracy

The autonomous multi-agent system produced a higher density of explicit factual claims per chapter. Claims were often clearly demarcated and enumerated, reflecting the operation of the internal fact-extraction agent. The human-in-the-loop system generated fewer discrete factual claims, instead embedding information within an explanatory narrative. This resulted in lower claim density but increased contextualization and hedging. Autonomous agents successfully identified potential inconsistencies and unsupported claims during critique phases; however, flagged issues were not always resolved and were occasionally reintroduced in subsequent revisions. Human oversight in the collaborative system corrected factual inaccuracies more consistently, though effectiveness depended on the human operator's domain knowledge. Neither system produced error-free output without human intervention.

4. Human Intervention Analysis

The frequency of human intervention differed markedly between systems. The human-in-the-loop system required frequent but lightweight interventions, primarily involving sequencing decisions, clarification of intent, and stylistic smoothing. The autonomous system required fewer interventions overall, but when interventions were needed, they were more substantial, often involving the correction of accumulated stylistic repetition or narrative disjunction. Interventions in the human-in-the-loop system were predominantly proactive, occurring during drafting. In contrast, interventions in the autonomous system were reactive, occurring after multiple passes by the agent.

5. Process Efficiency and Stability

Initial draft generation was faster in the autonomous multi-agent system due to the agents' parallel execution. However, revision cycles extended the total completion time as stylistic and coherence issues accumulated. The human-in-the-loop system progressed more slowly during drafting but required fewer large-scale rewrites.

Observed failure modes in the autonomous system included repetitive phrasing, stale transitions, and occasional deadlocks between critic and editor agents. Failures in the human-in-the-loop system were primarily related to human fatigue and coordination overhead.

VI. LIMITATIONS AND ETHICAL CONSIDERATIONS

1. Limitations

Language and Domain Scope: This study is restricted to English-language content and a single topical domain (personal finance). While this scope allows for controlled comparison across systems, it limits the generalizability of the findings. Language-specific features such as grammar, rhetorical conventions, and discourse structure may influence system performance. Similarly, different topical domains - such as legal, medical, or scientific writing - impose distinct demands for accuracy, evidence, and domain expertise. Therefore, the outcomes observed in this study may not be directly transferable to other languages or content areas.

Dataset and Prompt Constraints: The study uses a fixed set of prompts, outlines, and evaluation criteria, which may bias results toward the specific writing style or structure used. Alternative prompt strategies, varied editorial guidelines, or different outline formats could yield different performance outcomes.

Human Evaluation and Subjectivity: Although we employ standardized rubrics and inter-rater reliability checks, human evaluation remains inherently subjective. Assessments of coherence, narrative quality, and factual accuracy can vary based on the evaluator's background knowledge, writing expectations, and personal preferences. Furthermore, fatigue and evaluator bias may influence scoring, particularly when reviewing long-form documents. Future work should include larger evaluator pools, blinded assessments, and quantitative measures such as automated coherence or factuality metrics.

Temporal and Systemic Variability: LLMs and agent frameworks are continuously updated, and performance can vary across model versions, API changes, and underlying infrastructure. The results reported here reflect a specific system configuration at a particular time. Replication efforts may yield different outcomes as models evolve.

Limited Ground Truth and Reference Materials: In long-form nonfiction writing, there is often no single "correct" version of a document. The absence of a definitive ground truth makes objective comparison challenging. Although we use domain-appropriate reference materials and expert review where possible, the evaluation still relies on approximate measures of accuracy and relevance.

Reporting Granularity and Sample Size: This study compares a single run of each system rather than multiple trials, reporting implementation details at the level of roles, categories, and workflow structure rather than as verbatim prompts, model hyperparameters, or rater-level data. This scope was adopted to keep the manuscript focused on architectural comparison. It also means the findings should be read as descriptive observations from a comparative case study rather than as statistically validated or independently replicable results. Expanding the sample size and conducting quantitative evaluations along relevant dimensions constitute opportunities for future work.

2. Ethical Considerations

Transparency and Disclosure: AI-assisted writing blurs the boundary between human and machine authorship. This study emphasizes the importance of transparent disclosure regarding the extent of AI involvement. Failure to disclose AI contributions may mislead readers and violate academic integrity standards. For example, undisclosed AI authorship could compromise trust in scholarship and raise ethical issues regarding attribution.

Authorship and Ownership: Questions of intellectual property and authorship become more complex when multiple AI agents contribute to a manuscript. Determining ownership of generated content, especially when AI systems produce significant portions of text, is a legal and ethical challenge. Additionally, the role of human editors and supervisors must be clarified to avoid misrepresentation of contributions.

Bias and Fairness: LLMs can reproduce or amplify biases present in training data, including cultural stereotypes, gender bias, or socioeconomic assumptions. In domains like personal finance, biased recommendations can disproportionately affect vulnerable populations. Although multi-agent architectures may introduce checks, they do not inherently eliminate bias. Mitigating bias requires explicit intervention, diverse evaluation teams, and careful design of prompts and agent roles.

Misinformation and Hallucination Risks: The potential for hallucinated claims and fabricated facts is a major ethical concern. In high-stakes domains, misleading content may cause financial harm or reinforce false beliefs. Multi-agent systems can reduce but not fully eliminate hallucinations; human verification remains essential for responsible deployment. Additionally, systems must be designed to avoid flag uncertainty and avoid presenting unverified claims as fact.

Safety, Accountability, and Auditing: Autonomous multi-agent systems complicate accountability because multiple agents may contribute to the final output. Implementing and auditing guardrails - such as safety constraints, content filters, and fact-checking mechanisms - becomes challenging when agents operate asynchronously or in parallel. Transparent logging of agent decisions, traceability of content sources, and post-hoc auditability are necessary for responsible deployment, particularly in regulated domains.

Privacy and Data Security: The use of proprietary prompts, user data, or domain-specific reference materials raises privacy concerns. If agents access sensitive or personal information, there is a risk of unauthorized exposure or misuse. Researchers must ensure secure handling of data, adhere to privacy policies, and avoid using private datasets without consent.

VII. DISCUSSION AND CONCLUSION

The integrated results of this study reveal a consistent and measurable trade-off between autonomy and narrative quality in AI-assisted long-form authorship. The human-in-the-loop system produced prose that was more coherent, stylistically consistent, and reader-oriented. In contrast, the autonomous multi-agent system demonstrated superior structural rigor and internal monitoring capabilities. These outcomes align with prior research showing that human editors play a critical role in resolving deep semantic dependencies and maintaining discourse continuity, particularly over extended text spans [17], [18]. The superior narrative quality of the human-supervised system may be attributed to humans' capacity for *tacit knowledge integration* - the ability to make nuanced judgments about value, tone, and rhetorical effect that are not easily encoded in explicit planning heuristics [19]. Conversely, the autonomous multi-agent framework exhibited strengths in claim identification**,** structural adherence**,** and systematic verification but often struggled with the subtle rhetorical transitions that contribute to narrative flow. This observation is consistent with prior findings that agentic

architectures, while effective at modular subtask processing, do not inherently encode the pragmatic principles of storytelling and argumentation that human writers employ [20]. These findings suggest that narrative coherence remains a central unsolved challenge for autonomous multi-agent writing systems. While planner-driven architectures and subagent decomposition can enforce structural constraints and reduce hallucinations in modular tasks, they do not fully capture the tacit narrative judgments exercised by human collaborators. For example, human authors are more adept at maintaining thematic continuity across sections and anticipating reader expectations, both of which remain difficult to quantify in agent utility functions [21].

At the same time, the autonomous system's strengths - particularly in monitoring internal consistency, handling repetitive structural checks, and flagging logical gaps - indicate meaningful progress toward scalable AI authorship. These capabilities echo findings from recent work on self-refining language model pipelines, which demonstrate that iterative critique loops improve factuality and structural compliance, even without human intervention [22], [23]. Taken together, the results support hybrid models in which autonomous agents handle planning, drafting, and internal critique, while humans provide targeted oversight focused on narrative cohesion and ethical nuance. Such collaborative frameworks align with sociotechnical models of human–AI interaction that emphasize complementary strengths: humans contribute contextual judgment and ethical reasoning, while AI systems contribute computational scalability and process consistency [24], [25]. From a research perspective, these results underscore the importance of evaluating AI writing systems not only on output quality but also on process characteristics such as human labor distribution, failure recovery, and long-horizon stability. For instance, understanding how systems handle *drift* - the gradual erosion of narrative focus over successive generations - can inform design choices for persistent memory, agent re-evaluation triggers, and cross-agent feedback loops [26]. Furthermore, quantifying human effort in post-editing versus in-line guidance can inform cost–benefit analyses for deploying different levels of autonomy across writing tasks. Finally, while this study focuses on structured nonfiction, similar trade-off dynamics likely also occur in other genres, such as technical documentation, policy analysis, and educational materials. Future research should investigate how genre conventions and audience expectations interact with autonomous and hybrid authorship paradigms to shape narrative outcomes and ethical considerations.

1. Conclusion

This study provides a case study comparing two paradigms of AI-assisted long-form authorship, human-in-the-loop multi-persona collaboration and autonomous multi-agent orchestration, within a structured nonfiction writing task. The results indicate that multi-agent AI systems can generate structured, coherent long-form content, demonstrating meaningful progress in overcoming the limitations of single-agent LLM generation. However, the findings also suggest that human oversight remains critical for achieving high-quality narrative authorship, particularly in maintaining factual integrity, ensuring logical consistency, and preserving stylistic coherence across extended text. The comparative evaluation reveals trade-offs between autonomy and quality: autonomous systems offer greater efficiency and reduced human labor, but they are more prone to drift, hallucination, and inconsistencies in long-form contexts. In contrast, human-in-the-loop frameworks can better manage narrative continuity and factual accuracy but require greater time and expertise. As AI-assisted writing continues to evolve, understanding these trade-offs will be essential for responsible deployment. Future work should explore broader domains, multilingual settings, and more robust evaluation frameworks - including automated factuality checks and larger human evaluation cohorts. Moreover, research must continue to address ethical considerations such as transparency, authorship attribution, bias mitigation, and auditability to ensure that AI-generated texts are reliable, accountable, and aligned with human values.

ACKOWLEDGEMENT

All workflows were deterministic, given fixed prompts and random seeds. Agent definitions, execution scripts, and evaluation rubrics were preserved to enable replication. All generated content was labeled as AI-assisted, and no human subject's data were collected.

REFERENCES

[1]. P. A. Nylund, X. Ferràs-Hernández, and A. Brem, "A trust paradox may limit the application of AI-generated knowledge," *Res. Technol. Manage.*, vol. 66, no. 5, pp. 44–52, 2023, https://doi.org/10.1080/08956308.2023.2236475.

[2]. A. Bozkurt, "GenAI et al. cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI," *Open Praxis*, vol. 16, no. 1, pp. 1–10, 2024, https://doi.org/10.55982/openpraxis.16.1.654.

[3]. A. Adel and N. Alani, "Can generative AI reliably synthesize literature? Exploring hallucination issues in ChatGPT," *AI & Soc.*, vol. 40, pp. 6799–6812, Jun. 2025, https://doi.org/10.1007/s00146-025-02406-7.

[4]. Yi, Q., et al., "SCORE: Story Coherence and Retrieval Enhancement for AI Narratives," *arXiv.org*, 2025, https://doi.org/10.48550/arXiv.2503.23512.

[5]. S. Salehi, Y. Singh, K. K. Horst, Q. A. Hathaway, and B. J. Erickson, "Agentic AI and large language models in radiology: Opportunities and hallucination challenges," *Bioengineering*, vol. 12, no. 12, p. 1303, 2025, https://doi.org/10.3390/bioengineering12121303.

[6]. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H., "Retrieval-augmented generation for large language models: a survey," *arXiv.org*, 2024, https://arxiv.org/abs/2312.10997.

[7]. H. Xia, H. Peng, Y. Qi, X. Wang, B. Xu, L. Hou, and J. Li, "Storywriter: A multi-agent framework for long story generation," *arXiv.org*, Jun. 2025, https://arxiv.org/abs/2506.16445.

[8]. Y. Wu, Y. Bai, Z. Hu, J. Li, and R. Ka-Wei Lee, "Superwriter: reflection-driven long-form generation with large language models," *arXiv.org*, Jun. 2025, https://arxiv.org/abs/2506.04180.

[9]. Y. Yu et al., "An overview of large language models and a novel LLM-based cognitive architecture for solving open-ended problems," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 4, Nov. 2025, https://doi.org/10.3390/2504-4990-7-4-134.

[10]. M. Abou Ali, F. Dornaika, and J. Charafeddine, "Agentic AI: A comprehensive survey of architectures, applications, and future directions," *Artif. Intell. Rev.*, vol. 59, art. no. 11, 2026, https://doi.org/10.1007/s10462-025-11422-4.

[11]. T. Kwartler, M. Berman, and Aqrawi, "Good parenting is all you need - Multi-agentic LLM hallucination mitigation," *arXiv.org*, Oct. 2024, https://arxiv.org/abs/2410.14262.

[12]. L. Zhang and K. D. Ashley, "Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach," *arXiv.org*, Jun. 2025, https://arxiv.org/abs/2506.02992.

[13]. G. Papageorgiou, V. Sarlis, M. Maragoudakis, I. Magnisalis, and C. Tjortjis, "Evaluating faithfulness in agentic RAG systems for e-governance applications using LLM-based judging frameworks," *Big Data Cogn. Comput.*, vol. 9, p. 309, 2025, https://doi.org/10.3390/bdcc9120309.

[14]. A. Kasirzadeh and I. Gabriel, "In conversation with artificial intelligence: Aligning language models with human values," *Philos—Technol.*, vol. 36, no. 2, art. no. 27, 2023, https://doi.org/10.1007/s13347-023-00624-3.

[15]. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., "Training language models to follow instructions with human feedback," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 27730–27744, 2022, https://doi.org/10.52202/068431-2011.

[16]. J. Ding, L. M. Hughes, D. Berleant, A. W. Fulmer, and E. S. Wurtele, "PubMed Assistant: A biologist-friendly interface for enhanced PubMed search," *Bioinformatics*, vol. 22, no. 3, pp. 378–380, Feb. 2006, https://doi.org/10.1093/bioinformatics/bti821.
[17]. M. Lapata, "Evaluating content selection in summarization," *Lang. Linguist. Comput*., vol. 22, no. 3, pp. 409–432, 2012, https://doi.org/10.1093/llc/fqs023.
[18]. A. Paul and W. M. Fong, "Understanding narrative coherence in long form text generation," *Proc. Conf. Empirical Methods in Natural Language Process*., pp. 345–359, Oct. 2023, https://aclanthology.org/2023.emnlp-main.29.
[19]. S. Reddy, E. Demberg, and A. Menezes, "Tapping into tacit knowledge: challenges for computational models of pragmatics," *J. Artif. Intell. Res*., vol. 70, pp. 907–934, Apr. 2021, https://doi.org/10.1613/jair.1.12416.
[20]. Li and X. Zhang, "Agent-based modular generation: a survey of multi-agent architectures for text generation," *Artif. Intell. Review*, vol. 55, no. 2, pp. 1101–1132, Feb. 2022, https://doi.org/10.1007/s10462-021-10005-7.
[21]. X. Zhang and Y. Wang, "Toward narrative intelligence: evaluating narrative quality beyond perplexity," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 526–540, Dec. 2022, https://doi.org/10.1162/tacl_a_00436.
[22]. Z. Xu et al., "Self-critique and self-improvement in large language models," *Proc. Int. Conf. Learn. Representations (ICLR)*, May 2024, https://openreview.net/forum?id=xyz123.
[23]. L. Chen et al., "Iterative feedback loops for reducing hallucinations in LLMs," *arXiv.org*, Mar. 2025, https://arxiv.org/abs/2503.11289.
[24]. D. S. Weld and D. B. Smith, "Human-AI collaboration: current trends and open problems," *AI Magazine*, vol. 42, no. 3, pp. 15–32, Fall 2025, https://ojs.aaai.org/index.php/aimagazine/article/view/7775.
[25]. J. E. Fox, N. P. Johnson, and K. C. Davis, "Sociotechnical perspectives on AI-augmented writing," *J. Comput. Cult. Heritage*, vol. 15, no. 1, pp. 9:1–9:18, Feb. 2025, https://doi.org/10.1145/3447596.
[26]. B. Zhou and M. Marcus, "Measuring drift in generative text models over long horizons," *arXiv.org*, Nov. 2024, https://arxiv.org/abs/2411.04567.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Utilization of Honeypot Systems for Enhancing Cybersecurity in Smart Cities

Vladimir Sobeslav
Faculty of Management and Informatics
University of Hradec Králové
Hradec Králové, Czech Republic
vladimir.sobeslav@uhk.cz
ORCID:0000-00020917-5988

Daniel Kucera
Faculty of Management and Informatics
University of Hradec Králové
Hradec Králové, Czech Republic
daniel.kucera.2@uhk.cz
ORCID:0009-0008-3248-1132

Jan Pulpan
Faculty of Management and Informatics
University of Hradec Králové
Hradec Králové, Czech Republic
jan.pulpan@uhk.cz
ORCID:0009-0002-2626-0428

Abstract
The rapid evolution of cyber threats requires defense systems that can dynamically adapt to novel attacker behaviors. Honeypot systems—deceptive environments intentionally exposed to attract adversaries—provide valuable data for understanding attack techniques and tactics. Smart Cities rely on interconnected information technology (IT) and operational technology (OT) systems that support critical services such as transport, energy, and communication. As these infrastructures grow in complexity, their exposure to cyber threats increases, demanding continuous monitoring and trustworthy methods for collecting real-world attack data. This paper presents the design and deployment of a scalable honeypot system built on containerization principles to support cybersecurity monitoring and threat analysis. The containerized architecture ensures repeatability, isolation, and safe operation in both laboratory and production-like environments. The proposed framework provides a foundation for quantitative cybersecurity measurement and situational awareness. The framework can be adapted to Smart City infrastructures that integrate IT, OT, and IoT networks. Honeypot-based telemetry has been shown to support the assessment of network exposure, resilience metrics, and decision-making in trustworthy intelligent information systems. Recommendations are provided for extending the system toward Smart City-specific services and measurement indicators.

Keywords: Honeypot Systems, Containerization, Docker, Smart City Cybersecurity, IT/OT Convergence, Threat Measurement, Network Monitoring, Trustworthy Information Systems, Cybersecurity Monitoring, Situational Awareness

I. INTRODUCTION

Smart cities rely on interconnected Internet of Things (IoT), Industrial Internet of Things (IIoT), and cyber–physical system (CPS) components. These components support critical urban services such as transportation, energy distribution, water management, and public administration. The convergence enables automation but also expands the cyber-attack surface. It introduces constraints related to heterogeneity, real-time operation, and limited computational resources. Conventional cybersecurity mechanisms designed for homogeneous IT environments are insufficient. Specialized monitoring and deception-based security approaches are needed. Honeypot systems are an established defensive technique that deliberately exposes deceptive services to attract adversaries and observe their behavior. By interaction level, honeypots are classified as low-interaction, high-interaction, or hybrid, each trading off realism, operational risk, and resource demand. Deployed alongside firewalls and intrusion detection systems (IDS), honeypots provide labeled attack data and insight into attacker tactics, techniques, and procedures (TTPs). Recent research extends classical honeypot concepts toward cloud-based, edge-deployed, and virtualized environments, including digital twins. These approaches address advanced persistent threats (APTs) and ransomware targeting smart-city infrastructures. They emphasize early detection and safe

isolation, both essential where cyber incidents may have physical or societal consequences. This paper proposes a scalable, container-based honeypot framework for cybersecurity monitoring in smart-city environments that integrates IT, OT, and IoT networks. Lightweight containerization and modular service emulation enable repeatable deployment, controlled isolation, and adaptation to diverse smart-city domains. The framework is validated through real-world deployment and experimental observations, demonstrating its applicability for quantitative cybersecurity assessment and resilience evaluation in intelligent urban systems.

II. PROBLEM DEFINITION AND RELATED WORKS

1. Cybersecurity Challenges in Smart City Environments

The literature identifies several recurring vulnerability patterns in smart-city infrastructures, particularly related to IoT and edge devices [1]:

- Use of default or weak authentication credentials,
- exposed management interfaces reachable from the Internet,
- outdated or unpatched firmware and software,
- insecure or unencrypted communication protocols,
- limited monitoring and logging capabilities on resource-constrained devices.

Beyond individual device weaknesses, the high degree of interconnectivity between smart-city subsystems enables cascading attack scenarios. A compromise of a single component may provide attackers with pivot points for lateral movement, allowing them to gain persistent access to critical infrastructure, exfiltrate sensitive data, disrupt essential services, or coordinate attacks across multiple city domains [1]. Smart cities, therefore, attract a diverse range of threat actors, including nation-state attackers targeting critical infrastructure, cybercriminal groups motivated by financial gain (e.g., ransomware), hacktivists pursuing ideological objectives, and opportunistic attackers exploiting exposed devices for botnet recruitment and cryptomining.

2. Honeypot Systems: Concepts, Classification, and Research Evolution

Honeypots are deception-based security mechanisms that imitate legitimate systems, services, or devices to attract adversaries and monitor malicious activity. Unlike preventive security controls, honeypots primarily serve as intelligence-gathering tools, enabling detailed analysis of attacker behavior, tactics, techniques, and procedures (TTPs) within a controlled, isolated environment [2].

- Core characteristics common to honeypot systems include [3]:
- deception, making the system appear as a valuable target,
- isolation from production infrastructure to contain risk,
- detailed monitoring and logging,
- absence of legitimate user traffic, ensuring that any interaction is suspicious by definition.
- Based on interaction level, honeypots are commonly classified as [4]:

Low-interaction honeypots, which emulate limited service functionality with minimal resource requirements and low operational risk, are suitable for large-scale monitoring but provide relatively shallow intelligence.

High-interaction honeypots, which expose fully functional systems or services, enabling deep attacker engagement, malware collection, and complete attack-chain observation at the cost of higher complexity and containment risk. Hybrid honeypots, which combine low- and high-interaction elements, often dynamically scale interaction fidelity based on observed attacker behavior to balance scalability and intelligence depth. From a deployment perspective, honeypots may be used as production honeypots embedded within operational networks, research honeypots dedicated to threat intelligence collection, or distributed honeynets that provide broader coverage across geographically dispersed infrastructures [5]. Early honeypot research focused on traditional IT systems. The emergence of IoT and smart-city environments has since driven evolution toward protocol-specific emulation, cloud and edge deployment, machine learning–assisted analysis, and adaptive architectures such as digital twins and SDN-enabled honeypots [1],[6],[7].

3. Rationale for Honeypots in Smart City Security Monitoring

Smart-city infrastructures make honeypots a useful complement to traditional security mechanisms. Honeypots generate high-quality, labeled malicious traffic, which is essential for training and validating machine learning–based detection models. Since legitimate users do not interact with honeypots, all collected data can be treated as malicious. This removes the labeling ambiguity common in operational security logs [6],[1]. Honeypots also provide early warning capabilities by detecting reconnaissance, scanning, and initial exploitation attempts before attackers reach production systems. This is especially critical in smart cities, where cyber incidents may have direct physical or societal consequences and require rapid response [8]. Furthermore, honeypot observations enable systematic documentation of attack patterns and TTPs, including:

- exploitation techniques targeting IoT and industrial protocols,
- malware designed for resource-constrained devices,
- lateral movement strategies within interconnected city networks,
- automated scanning tools and attack frameworks.

By complementing traditional security controls, honeypots help detect novel or previously unseen attacks, reduce false positives in security operations centers, support forensic investigations, and enable resource-efficient monitoring of large and heterogeneous attack surfaces, particularly through low-interaction deployments suitable for IoT-rich environments.

III. HONEYPOT ARCHITECTURE FRAMEWORK FOR SMART CITIES

1. Smart City Honeypot Requirements

Smart City honeypots must operate in environments where cyberattacks on OT/IoT components can cause physical harm, service disruption, and threats to public safety – consequences absent from conventional IT incidents [9],[10],[11]. This chapter presents a honeypot [1] architecture framework designed for these environments, balancing threat detection with operational security.

Smart City honeypot deployments differ from conventional IT honeypots in five requirements driven by OT/IoT constraints [12],[13]:

Protocol diversity – support for industrial protocols (Modbus, DNP3, BACnet, IEC 61850) alongside IT and IoT protocols (MQTT, CoAP).

Distributed deployment – coverage across geographically dispersed infrastructure rather than centralized datacenter placement.

Resource efficiency – operation on resource-constrained network devices and edge gateways.

Extreme isolation – OT criticality demands zero tolerance for lateral movement from compromised honeypots to production infrastructure.

Domain-specific emulation – authentic simulation of Smart City infrastructure to avoid detection by adversaries familiar with these environments.

The proposed architecture addresses these requirements through a multi-tier deployment model, containerized implementation, and defense-in-depth isolation, detailed in the following sections.

Multi-Tier Deployment Architecture

We propose a multi-tier [14] architecture (Figure 1) that balances detection coverage, resource constraints, and risk. Smart Cities can deploy honeypot complexity [15] matched to available infrastructure and required intelligence depth.

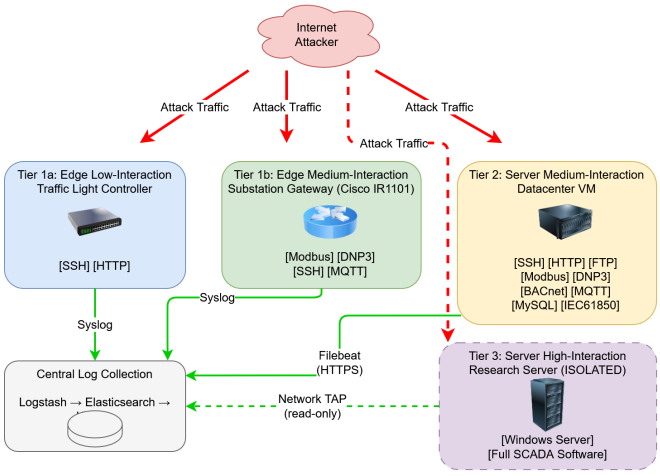


Figure 1. Multi-Tier Deployment in Smart City Infrastructure.

Tier 1a provides city-wide detection with minimal overhead on existing infrastructure. Tier 1b enables richer protocol interaction at critical points (substations, pump stations) where devices have sufficient resources. Tiers 2-3 follow conventional patterns, with Tier 3 reserved for research due to extreme isolation requirements.

Table 1. Multi-Tier Honeypot Deployment Model

Tier & Interaction	Platform Examples	Services	RAM	Purpose
1a: Edge Low	Basic switches, IoT gateways	1-3 emulators	256MB-1GB	Wide-area scanning detection
1b: Edge Medium	Cisco Cat 9K, industrial routers	4-8 handlers	2-4GB	Enhanced edge TTPs collection
2: Server Medium	Edge/datacenter VMs	10-15 stacks	4-8GB	Deep protocol intelligence
3: Server High	Dedicated research servers	Full OS + SCADA	16GB+	Advanced threat research

2. Core Architectural Components

Honeypot Layer. Containerized services run across all tiers with dedicated network namespaces. Multiple honeypots coexist on a single platform (e.g., a substation router hosting SSH, Modbus, and DNP3 containers simultaneously). Secure Data Collection Layer. Critical requirement: honeypots accept inbound attacker traffic but cannot initiate outbound connections to production networks. Log collection varies by platform: network devices use native syslog forwarding (UDP port 514), server containers use log shippers (Filebeat via HTTPS), and high-interaction VMs may employ network taps. Firewalls permit only honeypot-to-collector traffic on specific ports while blocking all honeypot-to-production communication (Figure 2).

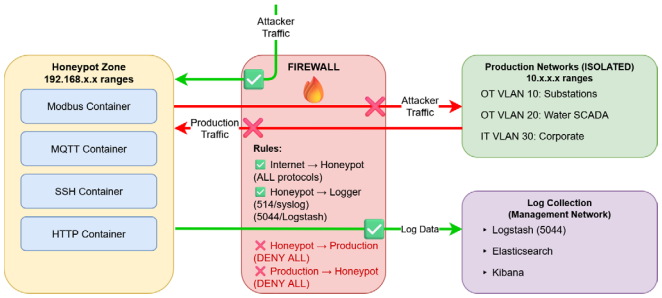


Figure 2. Data Collection Architecture and Traffic Flow Control.

Central Analysis Layer. Elastic Stack (Elasticsearch, Logstash, Kibana) receives distributed logs, enriches them with GeoIP and threat intelligence, normalizes formats, and provides visualization. Infrastructure resides on an isolated management network accessible only to authorized analysts.

Network Segmentation and Isolation

Defense in depth requires multiple isolation layers to prevent a compromise of the honeypot from affecting production [12].

Table 2. Isolation Layers by Platform.

Layer	Network Devices	Servers	Function
Container namespace	✓	✓	Process, filesystem, network isolation

Layer	Network Devices	Servers	Function
Virtual machine	✗	✓ (Tiers 2-3)	OS-level isolation
VLAN segmentation	✓	✓	Layer 2 separation
Firewall enforcement	✓	✓	Stateful filtering and logging
Physical separation	✗	✓ (Tier 3)	Ultimate isolation

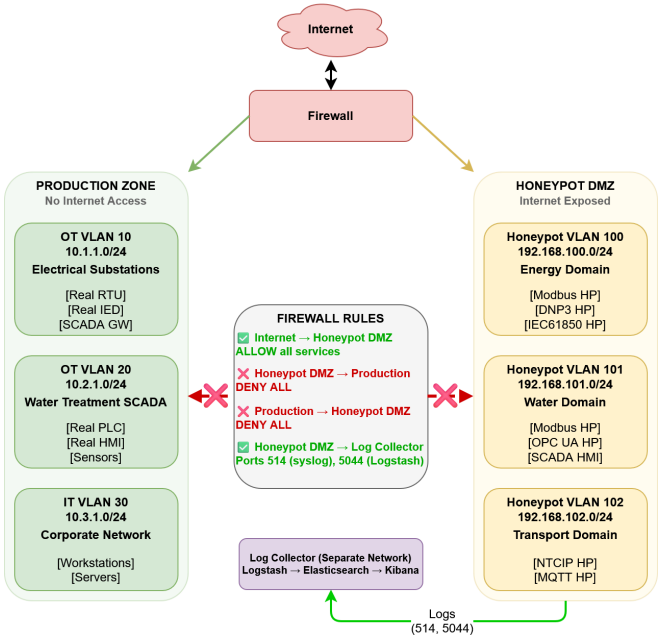


Figure 3. Network Segmentation Architecture.

VLAN segmentation (Figure 3) isolates honeypot networks (192.168.x.x/24 ranges) from production (10.x.x.x ranges). Firewall rules enforce: Internet→Honeypot ALLOW, Honeypot→Production DENY ALL, Honeypot→Log Collector ALLOW (specific ports only). Production networks have no routes to honeypot networks.

Containerization Strategy with Docker

Docker enables practical Smart City deployment for five reasons:

Portability – runs on x86 servers, ARM gateways, and network devices (Cisco IOx, IOS-XE).

Resource efficiency – Alpine-based images consume 30–50 MB.

Rapid deployment – pre-built images available (Cowrie, Conpot).

Isolation – namespace separation without VM overhead.

Orchestration – Docker Compose for single hosts, Kubernetes for city-wide deployment.

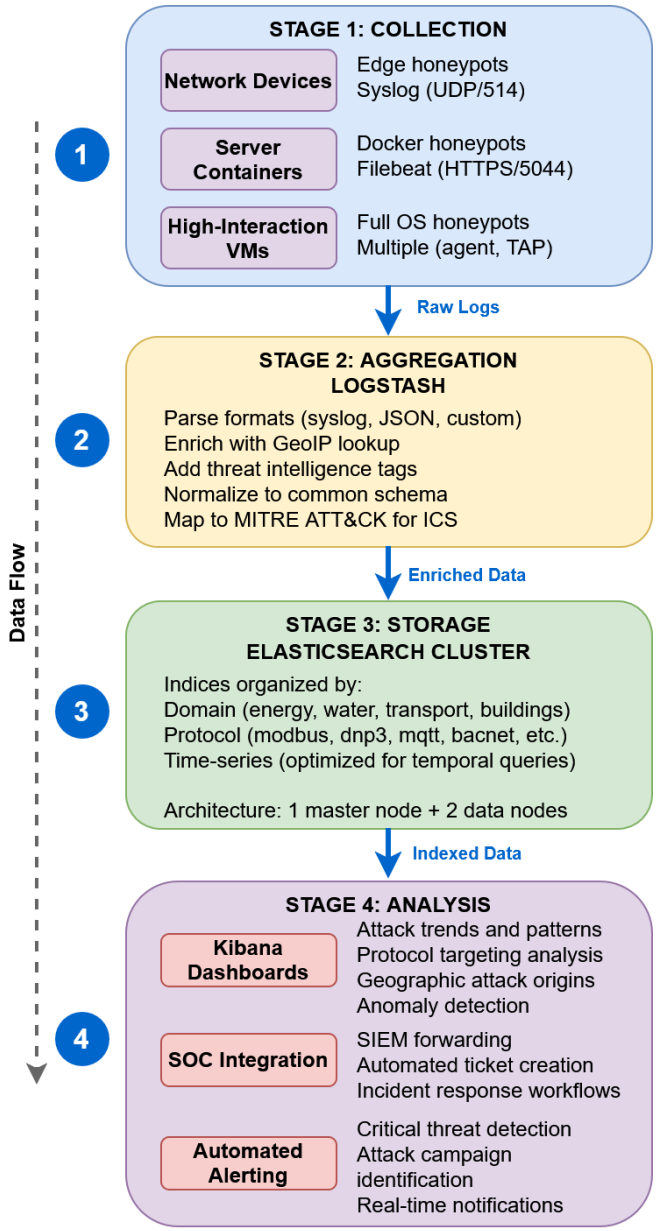


Figure 4. End-to-End Data Collection Pipeline.

Tier 1 edge uses 2–4 containers per device, selected by domain (energy: Modbus/DNP3, water: Modbus/OPC UA, transport: NTCIP/MQTT). Tier 2 servers deploy full stacks with 10–15 services plus log collection. Traffic controllers with Docker support host SSH, HTTP, and NTCIP containers using under 200 MB RAM total with minimal performance impact.

3. Data Collection Pipeline

All events normalize to a common JSON schema (Figure 4): temporal data; honeypot metadata (tier, location, device); Smart City context (domain, infrastructure); attack details (source, protocol, actions); and enrichment (geo, threat intel). Logstash enriches logs with GeoIP, threat intelligence, and MITRE ATT&CK [16],[17] mappings. Elasticsearch time-series indexing optimizes storage, while domain-specific indices enable targeted analysis. Kibana dashboards provide real-time visibility into attack trends, protocol patterns, geographic origins, and anomalies.

IV. SMART CITY DOMAIN APPLICATIONS AND DEPLOYMENT SCENARIOS

1. Smart City Domains and Threat Landscape

Smart Cities comprise multiple operational domains with distinct infrastructure characteristics, protocol ecosystems [18], and threat profiles. Effective honeypot deployment requires understanding domain-specific properties, as attack patterns and adversary motivations vary significantly across sectors.

Table 3. Smart City Critical Domains.

Domain	Infrastructure	Protocols	Threats & Impact	Priority
Energy	Grids, substations, smart meters	Modbus, DNP3, IEC 61850, IEC 60870-5	Protocol exploits, malware (Industroyer, Triton) → Blackouts, equipment damage	●□●□●□●□●□
Transport	Traffic control, transit, EV charging	NTCIP, OCPP, C-ITS, GTFS-rt	Signal manipulation, GPS spoofing, DoS → Traffic chaos, accidents	●□●□●□●□
Water	Treatment plants, distribution, and pumps	Modbus, DNP3, Profibus, OPC UA	Register manipulation, sensor spoofing → Contamination, flooding	●□●□●□●□●□
Buildings	BMS, HVAC, lighting, access control	BACnet, Modbus, KNX, LonWorks	BACnet exploits, HVAC manipulation → Discomfort, energy theft, breach	●□●□●□

Table 3 shows four critical Smart City domains, prioritized by the severity of consequences. Energy and water receive the highest priority due to the potential for physical harm and essential service disruption. Transportation warrants high priority, given public safety implications. Buildings present medium priority with a significant but less catastrophic impact. This prioritization guides resource allocation when budget or infrastructure constraints

prevent full coverage. Lower-priority domains (public services, environmental monitoring) follow similar deployment patterns but receive reduced coverage.

2. Protocol-to-Domain Mapping and Honeypot Placement Strategy

Effective honeypot deployment requires matching protocol emulation to domain-specific attack surfaces. Different Smart City domains use distinct protocol ecosystems—adversaries targeting specific infrastructure probes for the corresponding protocols during reconnaissance.

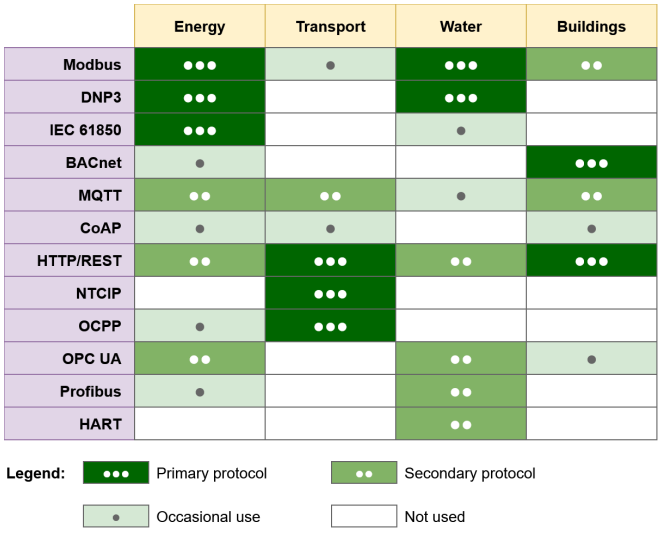


Figure 5. Protocol Distribution Across Smart City Domains.

Protocol distribution analysis (Figure 5) reveals that certain protocols span multiple domains (Modbus appears in energy, water, and buildings) while others remain domain-specific (NTCIP exclusively in transportation, IEC 61850 primarily in energy substations). This distribution informs honeypot selection: multi-domain protocols provide broad threat visibility but require domain-specific configuration to achieve authentic emulation.

Table 4. Honeypot Deployment Strategy.

Domain	Edge (Tier 1)	Server (Tier 2)	Density
Energy	Modbus, DNP3 on substation gateways	IEC 61850 devices, SCADA HMI web	1 edge/substation, 1 server/region
Transport	NTCIP on traffic controllers	Traffic mgmt web, OCPP central system	1 edge/50-100 intersections, 1 server/city
Water	Modbus on pump station PLCs	SCADA HMI, OPC UA server, Modbus master	1 edge/facility, 1 server/utility
Buildings	BACnet on BMS gateways	Building automation server, HVAC controller	1 edge/large building, 1 server/campus

Edge deployment positions lightweight containers at domain entry points (substations, pump stations, traffic cabinets) where these protocols are expected. This provides early warning during attacker reconnaissance, before production systems are reached. Server deployments in DMZ segments emulate centralized control systems (SCADA HMIs, traffic management centers, building automation servers) as high-value targets. Deployment density balances detection coverage against infrastructure and maintenance overhead. Critical domains (energy, water) justify higher edge density.

3. Detailed Deployment Scenarios

Scenario A: Energy Sector — Electrical Substation Protection

Infrastructure Context. Electrical substations contain Remote Terminal Units (RTUs), Intelligent Electronic Devices (IEDs), and SCADA gateways controlling circuit breakers, transformers, and protective relays. Modern substations communicate using IEC 60870-5-104 and DNP3 for SCADA telemetry, alongside IEC 61850 for substation automation, all of which have documented vulnerabilities and available attack tools.

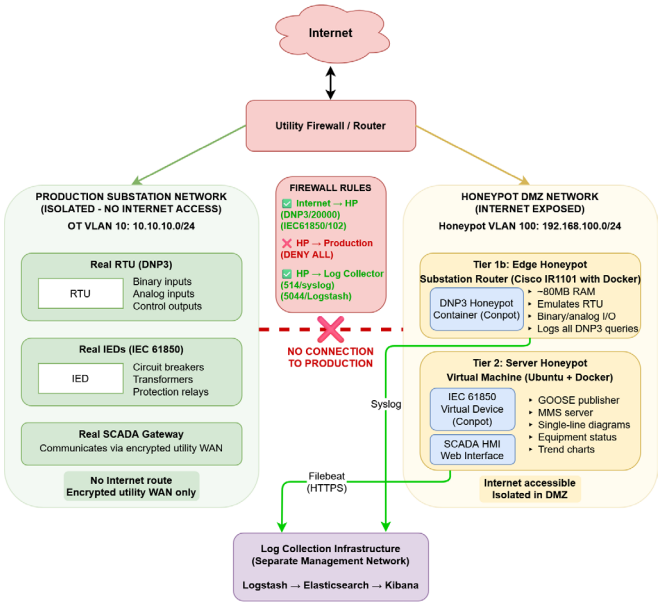


Figure 6. Electrical Substation Honeypot Deployment Architecture.

Deployment (Figure 6). Edge deployment uses an industrial router (Cisco IR1101) supporting containers via Cisco IOx application hosting. A Conpot IEC 60870-5-104 honeypot container (~80MB RAM) emulates a substation outstation with single-point information (circuit breaker status), measured values (voltage, current, power), and command points, responding to interrogation requests with synthetic data and logging control attempts. Server deployment operates in a utility data center DMZ on an Ubuntu VM hosting Conpot, configured with multiple IEC 60870-5-104 outstation profiles representing aggregated substation views, plus a custom web-based SCADA HMI displaying single-line diagrams and equipment status. Production equipment resides on private VLAN 10 (10.10.10.0/24) with no Internet route. Honeypots reside on VLAN 100 (192.168.100.0/24) in the DMZ, with Internet exposure. Firewall rules prohibit honeypot-to-production traffic and permit only honeypot-to-log-collector communication.

Expected Intelligence. Port scans targeting IEC 60870-5-104 (TCP port 2404) were detected before attackers located production systems. Automated scanners perform basic enumeration, while sophisticated adversaries issue protocol-specific queries that indicate expertise. Exploitation attempts capture malformed protocol messages targeting known vulnerabilities. Malware samples may include ICS-specific threats, such as Industroyer variants. This intelligence lets utility teams identify active targeting, obtain indicators of compromise for defensive tuning, and share threat intelligence with E-ISAC.

Scenario B: Water Sector — Treatment Plant SCADA Protection

Infrastructure Context. Water treatment facilities employ SCADA systems controlling chemical dosing pumps, filtration, disinfection, and distribution pumps, predominantly using Modbus TCP. The 2023–2024 CyberAv3ngers campaign targeting exposed Unitronics PLCs in U.S. water utilities [19] demonstrates real-world targeting of municipal water SCADA via exposed remote-access interfaces.

Deployment. Tier 2 server-based honeypot in utility DMZ uses Conpot Modbus with a custom template. Holding registers (1000-1099) emulate actuator setpoints: pump speeds (0-100%), valve positions, and chemical dosing rates. Input registers (2000-2099) emulate sensor readings, including flow rates, chemical concentrations (chlorine, pH, turbidity), tank levels, and pressure. The companion web-based SCADA HMI provides HTTP/HTTPS access with process diagrams, real-time trends, and operator controls, and implements basic authentication with common default credentials to entice attackers—expected Intelligence. Reconnaissance appears as Modbus read queries (function codes 0x03 and 0x04) that scan the register space to enumerate capabilities. Sabotage attempts use write commands (0x06, 0x10) to modify setpoints. HMI honeypot captures credential brute-forcing, SQL injection, and session hijacking. Attacker dwell time distinguishes opportunistic scanning from deliberate reconnaissance, suggesting the presence of targeted campaigns. Protocol command sequences reveal attacker familiarity, providing early warning of advanced persistent threat activity. Expected Threat Intelligence Outcomes and Community Value. Honeypot deployments generate intelligence categories with operational value that varies by tier and interaction level. Proactive defense detects campaigns targeting Smart City infrastructure before adversaries reach production systems. Security teams gain time to patch vulnerabilities, tune monitoring, or implement mitigations.

Table 5. Intelligence by Tier.

Intelligence Type	T1-Low	T1-Med	T2	T3
Scanning patterns	●●●	●●●	●●	●
Credentials tried	●●	●●●	●●●	●●●
Protocol commands	–	●●	●●●	●●●
Malware samples	–	●	●●	●●●
Dwell time/exploration.	–	●	●●	●●●
Lateral movement	–	–	●	●●●

Legend: ●●● = High value, ●● = Medium, ● = Low, – = Not captured

Tier 1 honeypots capture high-volume scanning activity, providing quantitative data on attack frequency, source distribution, and protocol-targeting trends. Tier 2 and 3 enable qualitative analysis: attacker methodologies, indicators of compromise, and threat-actor attribution.

Captured attack patterns inform intrusion detection rules, firewall policies, and SIEM correlation tailored to actual threats rather than generic signatures. The documented monitoring evidence supports sector-specific

regulation (e.g., NERC CIP for energy and EPA requirements for water). Beyond the individual organization, Smart City honeypot deployments contribute to community threat intelligence through ISACs (E-ISAC, WaterISAC, MS-ISAC) and national CERTs. Standardized sharing formats (STIX/TAXII) enable automated intelligence exchange. Mapping observed behaviors to MITRE ATT&CK for ICS lets smaller municipalities benefit from intelligence gathered by larger Smart Cities.

V. CONCLUSION

This paper addressed the growing cybersecurity challenges faced by smart-city infrastructures arising from the convergence of IT, OT, and IoT systems and the resulting expansion of the attack surface. Traditional preventive security mechanisms alone provide limited visibility into real attacker behavior, particularly during reconnaissance and early attack phases. To mitigate this gap, the paper proposed a scalable, container-based honeypot framework designed to support continuous cybersecurity monitoring and threat intelligence collection in smart-city environments. The main contribution is a modular honeypot architecture based on lightweight containerization. The framework enables repeatable deployment, strong isolation, and flexible service emulation in both laboratory and production-like environments. Experimental deployment showed that even a limited set of emulated services attracts significant malicious activity. This confirms honeypots as a source of real-world attack telemetry. The collected data provide quantitative insights into attack frequency, targeted services, credential abuse, and attacker behavior, supporting situational awareness and security assessment. The paper also presented a structured adaptation of the framework to smart-city environments. Domain-specific requirements, a multi-tier deployment model, and protocol-aware placement strategies show how honeypots can be integrated into complex urban infrastructures while respecting strict isolation. The analysis shows that honeypots complement traditional security controls by enabling early warning, improving forensic capabilities, and supplying labeled datasets for machine–learning–based detection. Several limitations should be acknowledged. The deployment focuses on IT-oriented services, and the results cannot be directly generalized to all OT and IoT protocols without further validation. Long-term deployment across multiple smart-city domains is required to fully assess attacker adaptation and the resilience of the deception mechanisms. Future work will focus on OT and IoT-specific honeypot modules, integration of collected telemetry with digital-twin-based cybersecurity platforms, and long-term distributed deployments across smart-city domains. These steps further strengthen the use of honeypots as a measurement and decision-support tool for resilience evaluation in intelligent urban systems.

REFERENCES

[1]. J. Franco, A. Aris, B. Canberk, and A. S. Uluagac, “A survey of honeypots and honeynets for Internet of Things, Industrial Internet of Things, and cyber-physical systems,” IEEE Commun. Surveys Tuts., vol. 23, no. 4, pp. 2351–2383, 2021, doi: 10.1109/COMST.2021.3106669.

[2]. S. Lanz et al., “Optimizing Internet of Things honeypots with machine learning: A review,” Applied Sciences, vol. 15, no. 10, Art. no. 5251, 2025, doi: 10.3390/app15105251.

[3]. L. Spitzner, Honeypots: Tracking Hackers. Boston, MA, USA: Addison-Wesley Professional, 2002.

[4]. I. Mokube and M. Adams, “Honeypots: Concepts, approaches, and challenges,” in Proc. 45th Annu. ACM Southeast Reg. Conf. (ACMSE), 2007, pp. 321–326, doi: 10.1145/1233341.1233399.

[5]. M. L. Bringer, C. A. Chelmecki, and H. Fujinoki, “A survey: Recent advances and future trends in honeypot research,” Int. J. Comput. Netw. Inf. Secur., vol. 4, no. 10, pp. 63–75, 2012, doi: 10.5815/ijcnis.2012.10.07.

[6]. Y. Ahmed, K. Beyioku, and M. Yousefi, “Securing smart cities through machine learning: A honeypot-driven approach to attack detection in Internet of Things ecosystems,” IET Smart Cities, vol. 6, no. 3, pp. 180–198, 2024, doi: 10.1049/smc2.12084.

[7]. M. Houichi, F. Jaidi, and A. Bouhoula, "Cyber security within smart cities: A comprehensive study and a novel intrusion detection-based approach," Comput. Mater. Continua, vol. 81, no. 1, pp. 393–441, 2024, doi: 10.32604/cmc. 2024.054007.

[8]. M. Choudhary, Mastering Honeypots: Art of Deception for Cybersecurity Defense. BPB Publications, 2024.

[9]. R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," IEEE Secur. Privacy, vol. 9, no. 3, pp. 49–51, 2011, doi: 10.1109/MSP.2011.67.

[10]. A. Cherepanov, "Win32/Industroyer: A new threat for industrial control systems," ESET, Tech. Rep., 2017.

[11]. A. Carcano, M. Krotofil, and Y. Dragoni, "TRITON: How it disrupted safety systems and changed the threat landscape of industrial control systems, forever," Black Hat USA Briefings, White Paper, 2018.

[12]. K. Stouffer et al., "Guide to operational technology (OT) security," NIST, Gaithersburg, MD, USA, Special Publication 800-82 Rev. 3, 2023, doi: 10.6028/NIST.SP.800-82r3.

[13]. Cybersecurity and Infrastructure Security Agency (CISA), "Cybersecurity best practices for smart cities," 2023. [Online]. Available: https://www.cisa.gov/sites/default/files/2023-04/cybersecurity-best-practices-for-smart-cities_508.pdf

[14]. M. Dodson, A. R. Beresford, and M. Vingaard, "Using global honeypot networks to detect targeted ICS attacks," in Proc. 12th Int. Conf. Cyber Conflict (CyCon), 2020, pp. 275–291.

[15]. A. V. Serbanescu, S. Obermeier, and D. Y. Yu, "A scalable honeynet architecture for industrial control systems," in ICETE 2015, ser. CCIS, vol. 585. Cham, Switzerland: Springer, 2016, pp. 179–200, doi: 10.1007/978-3-319-30222-5_9.

[16]. O. Alexander, M. Belisle, and J. Steele, "MITRE ATT&CK for industrial control systems: Design and philosophy," MITRE, Tech. Rep., 2020. [Online]. Available: https://attack.mitre.org/docs/ATTACK_for_ICS_Philosophy_March_2020.pdf

[17]. MITRE Corporation, "Techniques – ICS | MITRE ATT&CK." Accessed: Nov. 12, 2025. [Online]. Available: https://attack.mitre.org/techniques/ics/

[18]. G. Clarke, D. Reynders, and E. Wright, Practical Modern SCADA Protocols: DNP3, 60870.5 and Related Systems. Oxford, U.K.: Newnes, 2004.

[19]. Cybersecurity and Infrastructure Security Agency (CISA) et al., "IRGC-affiliated cyber actors exploit PLCs in multiple sectors, including U.S. water and wastewater systems facilities," Joint Cybersecurity Advisory AA23-335A, 2023. [Online]. Available: https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-335a

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Systematic Analysis of Honeypot Deception Techniques: State-of-the-Art and Empirical Validation

Vladimir Sobeslav
Faculty of Management and Informatics
University of Hradec Králové
Hradec Králové, Czech Republic
vladimir.sobeslav@uhk.cz
ORCID:0000-00020917-5988

Daniel Kucera
Faculty of Management and Informatics
University of Hradec Králové
Hradec Králové, Czech Republic
daniel.kucera.2@uhk.cz
ORCID: 0009-0008-3248-1132

Jan Pulpan
Faculty of Management and Informatics
University of Hradec Králové
Hradec Králové, Czech Republic
jan.pulpan@uhk.cz
ORCID:0009-0002-2626-0428

Abstract
Honeypot systems rely on deception realism to attract and analyze malicious actors, yet the gap between implementation sophistication and attacker detection capability continues to widen. This paper analyzes honeypot deception techniques spanning 20 years of research and is validated through an empirical deployment that captured 75,598 attacks across five protocols over eight days. We identify seven deception categories and assess their maturity through literature analysis and real-world attack data. The deployment showed that internet-exposed honeypots faced their first attack within three minutes, with SSH accounting for 96.12% of attacker traffic across 10+ countries. The analysis identifies response-timing manipulation as the most severe research gap, with only three significant works published despite its clear practical relevance. Cross-protocol consistency and OPC UA support emerge as additional high-priority gaps. We contribute a seven-category taxonomy, empirical validation of attack patterns, and a priority-ranked research roadmap.

Keywords: Honeypot, Deception techniques, Attack analysis, Response timing, Protocol emulation

I. INTRODUCTION

Within three minutes of deploying our multi-protocol honeypot system on July 21, 2025, the first attack against the exposed MySQL service began. Over eight days, we recorded 75,598 attacks targeting five network services from attackers in 10+ countries. This aggressive discovery pattern exemplifies the threat landscape honeypot systems must navigate while maintaining deceptive realism. Honeypot technology has evolved considerably since Provos introduced Honeyd in 2004 [1], moving from static service emulation toward fingerprinting evasion and, more recently, AI-driven adaptive systems. Yet fundamental challenges persist. Sophisticated attackers employ multi-vector detection probing service banners, response timing characteristics, filesystem artifacts, and cross-protocol consistency. A large-scale measurement recording over 400 million SSH brute-force attempts from more than four million unique source addresses found that attackers vary their SSH client version strings to evade signature-based detection [2]. Contemporary deployments face a paradox: research shows that 72% of Cowrie and Kippo deployments remain detectable through default-configuration artifacts, including predictable MAC addresses and uniform timestamps [3]. In contrast, the research community has invested heavily in banner manipulation and protocol emulation but has left response-timing manipulation severely underexplored. Attackers can bypass sophisticated protocol emulation by measuring response latencies that no real system could achieve. This paper addresses the capabilities and limitations of honeypot deception through three research questions, including: (RQ1) What is the current state of the art in deception techniques? (RQ2) How do real-world attack patterns validate existing approaches? (RQ3) What critical research gaps require attention? We conducted a systematic literature review spanning 2004-2025, deployed a multi-protocol honeypot capturing nearly 76,000 attacks, and synthesized findings to identify high-priority research directions. Our contributions are fourfold: (1) a seven-category taxonomy of deception techniques with maturity assessments, (2) empirical

validation through an eight-day deployment characterizing attack patterns including temporal dynamics, geographic distribution, and command execution sequences, (3) a gap analysis identifying response timing manipulation as the most severe deficiency alongside cross-protocol consistency and OPC UA support, (4) a synthesis framework connecting attack behaviors to deception requirements for evidence-based prioritization.

II. RESEARCH METHODOLOGY

Our approach combines systematic literature analysis with empirical validation of deception behavior under live attack traffic.

1. Systematic Literature Review

We conducted a systematic literature review covering 2004-2025, searching Scopus, Web of Science, and Semantic Scholar with keywords such as "honeypot," "deception," "fingerprinting," and "evasion." Inclusion criteria required a peer-reviewed publication, an explicit focus on deception, and technical implementation or empirical evaluation. This produced approximately 120 highly relevant papers spanning IT infrastructure, ICS/SCADA, IoT [4], and cloud deployments. We identified seven deception categories: banner manipulation, protocol emulation, interactive environment simulation, response timing manipulation, filesystem/artifact deception, AI-driven adaptive deception, and cross-protocol consistency. Each category was assessed across research volume, implementation maturity, empirical validation, and tool availability.

2. Empirical Deployment Study

We deployed a multi-protocol honeypot from July 21-29, 2025 (8 days, 3 hours, 21 minutes) using Docker containers that hosted SSH (Cowrie with a custom configuration), MySQL, DNS, FTP, and SMTP on standard ports, and exposed them via a public IP address. No active marketing occurred – all traffic represents organic discovery. We captured source IPs (geolocated), credentials, commands, temporal patterns, and session characteristics. For SSH, complete session transcripts enabled detailed command analysis. The deployment adhered to ethical guidelines, collecting no personal data beyond attack metadata.

Table 1. Research Methodology Overview.

Phase	Activity	Details
Literature Search	Database queries	Scopus, WoS, Semantic Scholar; 2004-2025
Screening	Inclusion criteria	Deception-focused, peer-reviewed, implementation
Categorization	Taxonomy	Seven deception categories
Empirical	Deployment	8 days, 5 protocols, 75,598 attacks
Synthesis	Gap analysis	Maturity vs. practical requirements

III. STATE-OF-THE-ART ANALYSIS

We analyzed honeypot deception techniques across the seven categories. Maturity varies substantially: some areas show 20+ years of development, while others remain severely underexplored. Figure 1 positions the categories by research maturity against detection risk, exposing the divergence between where research effort has accumulated and where attacker detection capability concentrates.

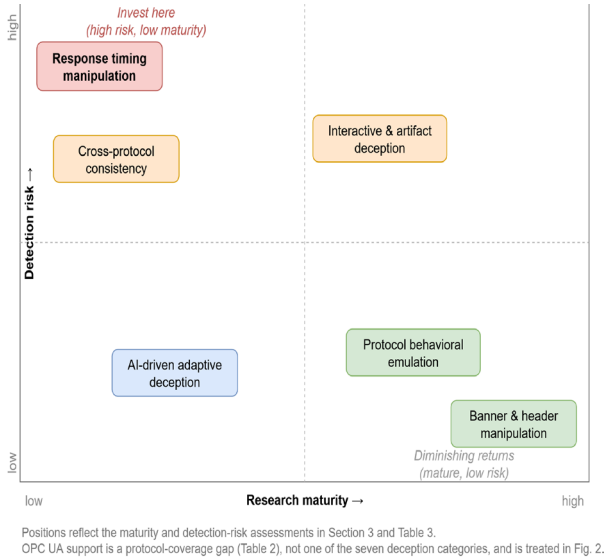


Figure 1. Deception categories, organized by research maturity (Section III), against detection risk. Response timing and cross-protocol consistency occupy the upper-left region, combining high detection risk with minimal research, while banner manipulation and protocol emulation sit in the mature, lower-risk region of diminishing returns. The detection-risk and maturity ratings are those consolidated in Table 3 (Section V-C). OPC UA support is a protocol-coverage gap (Table 2), not one of the seven deception categories.

1. Banner and Header Manipulation

Banner and header manipulation represents the most mature deception category with over two decades of continuous research since Honeyd's introduction in 2004 [1]. The foundational approach modifies service identification strings transmitted during initial protocol negotiation to mask honeypot characteristics and emulate target systems. Honeyd pioneered comprehensive banner manipulation, including TCP/IP stack fingerprinting evasion through precise control of packet characteristics such as TTL values, window sizes, and TCP options that operating system fingerprinting tools exploit. The SSH protocol has received particular research attention given its ubiquity as an attack vector. Early implementations employed static OpenSSH version strings, but the detection arms race has intensified. A multi-year measurement of large-scale SSH brute-force activity recorded over 400 million attempts from more than four million unique source addresses and observed 171 distinct SSH client version strings used to evade signature-based detection [2]. HTTP response header manipulation appears across multiple honeypot implementations, including Cowrie, Dionaea, and Chameleon, with Chameleon dynamically generating responses that match the emulated target device [5]. However, banner manipulation alone is insufficient to deceive sophisticated attackers. Modern reconnaissance tools perform version detection probing that extends well beyond initial banners, examining behavioral responses to protocol-specific queries. The research assessment shows extensive validation through production deployments, multiple mature open-source implementations, and well-understood techniques, though the ongoing arms race indicates continued evolution.

2. Protocol Behavioral Emulation

Protocol behavioral emulation extends beyond surface-level banner strings to implement realistic protocol state machines, command responses, and interaction patterns. Maturity varies dramatically across protocol domains, with traditional IT protocols achieving moderate-to-high fidelity while emerging protocols remain underserved. SSH protocol emulation through Cowrie is among the most sophisticated implementations, providing full SSH protocol support, including SFTP file transfer, SCP secure copy, and port forwarding [6]. The system implements an interactive shell simulation with comprehensive command-response libraries that handle common reconnaissance patterns. FTP protocol implementations must carefully simulate state machine transitions across LOGIN, CWD, LIST, and RETR/STOR sequences that attackers expect. SMTP honeypots face the delicate challenge of appearing realistic as mail relays without actually accepting open relay abuse. ICS and SCADA protocol emulation demonstrates rapidly maturing research aligned with the growing targeting of industrial control systems. Modbus protocol receives universal support across major ICS honeypots, including Conpot [7] and SecuriOT, with the latter capturing 54,682 Modbus packets during deployment validation [8]. The protocol's well-defined register maps and function codes enable high-fidelity emulation when properly configured. S7comm protocol targeting Siemens PLCs achieved breakthrough sophistication with HoneyPLC, which successfully deceived Nmap active fingerprinting, Shodan search indexing, and even specialized engineering software through careful implementation using the Snap7 protocol library [9]. DNP3 traffic grew from 5.1% to 22% of ICS scanning between 2015 and 2019, demonstrating increasing attacker interest [8]. Database protocols, including MySQL and PostgreSQL, exhibit limited sophistication in most honeypot implementations despite active targeting by attackers. Most MySQL honeypots perform basic authentication simulation without realistic query processing or database structure emulation that advanced attackers might probe. A critical gap emerges in OPC UA protocol support: our literature review identified only minimal dedicated OPC UA honeypot research [10],[11], despite OPC UA's status as the modern standard that replaces legacy ICS protocols. This represents a substantial vulnerability in industrial honeypot capabilities, given OPC UA's security-focused design and rapid adoption across industries. Table 2 summarizes the maturity of protocol emulation across key protocols.

Table 2. Protocol Emulation Maturity Assessment.

Protocol	Research Works	Production Tools	Fidelity	Gap Priority
SSH	20+	Cowrie, Kippo	High	Low
Modbus	15+	Conpot, SecuriOT	High	Low
S7comm	10+	HoneyPLC, Conpot	High	Low
DNP3	5+	Conpot	Medium	Medium
OPC UA	2	None	Low	High
MySQL	3	Dionaea	Low	High

3. Interactive Environment and Artifact Deception

An interactive environment simulation creates believable operating system contexts, including filesystem structures, process listings, system users, and command responses, that withstand reconnaissance probing. Early approaches employed static configurations: Kippo implemented fixed Debian-based filesystem hierarchies that proved effective against automated scanners but vulnerable to manual inspection by experienced attackers who recognized identical structures across deployments.

Modern implementations, including Cowrie, provide customizable filesystem definitions via configurable fs. pickle files, enabling deployment-specific variations. However, practical adoption lags behind technical capability: research analyzing real-world Cowrie and Kippo deployments found that 72% utilize default configurations readily identifiable through characteristic artifacts [3]. These detection vectors include predictable MAC addresses such as the default "88111222" sequence, standard hostname patterns, uniform filesystem structures, and suspicious timestamp uniformity where all files share identical modification dates corresponding to deployment timing rather than realistic system evolution. Filesystem realism requires multiple consistency dimensions. File counts must appear neither suspiciously sparse nor unrealistically dense. Modification timestamps should span months or years, reflecting genuine system history rather than a concentrated clustering around deployment dates. User home directories demand realistic personalization artifacts, including configuration files, shell histories, and application data appropriate to simulated user roles. System log files must contain realistic content, including authentication events, service startups, and cron job executions, with proper log rotation that demonstrates long-term operation. Command response simulation faces particular challenges given the attacker's use of sophisticated reconnaissance sequences. Common probing patterns include uname system information queries with various flag combinations to cross-check consistency, /proc/uptime filesystem access testing both file existence and content realism, complex shell pipelines using cut, awk, and arithmetic operators to verify shell capability, and deliberate write-test sequences such as echo test > /tmp/file && cat /tmp/file && rm /tmp/file explicitly designed to detect nonpersistent or read-only filesystems that betray honeypot implementations. Sophisticated attackers maintain IP blocklists of confirmed honeypot addresses discovered through artifact matching. The research community has thoroughly documented detection artifacts and proposed mitigation approaches through randomization, borrowing from real systems, and continuous evolution, yet the high default-configuration rate indicates an operational gap rather than a technical knowledge gap requiring improved tooling.

4. Response Timing Manipulation

Response timing manipulation represents the most severe research gap, with only three significant works over 20 years, despite its clear practical relevance. This addresses temporal characteristics that sophisticated attackers measure to detect artificial environments exhibiting impossible response patterns. Detection vectors fall into four areas:

- instantaneous responses betraying absent processing overhead, where real systems incur CPU scheduling, disk I/O, and network stack delays
- geographic latency absence when honeypots respond uniformly regardless of attacker location (Asian attackers expect 200-300ms transatlantic RTT, honeypots respond in <1ms)
- load-dependent timing variations are absent from honeypots that maintain constant response times
- protocol-specific timing when SSH key exchanges or MySQL queries complete impossibly quickly

Limited existing research provides minimal coverage. Endless SSH tarpit (2019) deliberately implements slow banner transmission, prolonging connections beyond 10 minutes to cause resource exhaustion rather than realistic deception [12]. Variable delay experiments propose random insertion without systematic frameworks. Recent works (2023-2024) acknowledge timing problems but offer no comprehensive solutions. Nevertheless, universal applicability combined with minimal research attention establishes this as the highest-priority gap requiring immediate investment. Section V-A develops the resulting research agenda.

5. AI-Driven Adaptive Deception

AI-driven deception represents the newest category experiencing dramatic growth (2023-2025) as LLMs and reinforcement learning transition from theoretical to experimental. LLM integration enables dynamic command-

response generation, handling unexpected commands, generating on-demand filesystem content, and adapting to attacker sophistication [13]. However, critical weaknesses persist: prompt injection vulnerabilities that enable attacker manipulation, model hallucinations that produce unrealistic responses, computational costs that exceed those of traditional approaches, limited validation with no long-term deployment studies, and zero adversarial robustness research on AI-powered attacker detection. RL approaches show theoretical promise: Wagener et al. employed Q-learning for adaptive behavior, including command blocking and attacker engagement [14], while Dowling et al. demonstrated that RL can overcome honeypot detection [15]. Assessment: nascent stage despite rapid growth, very limited production deployments, completely unexplored adversarial robustness, unresolved security vulnerabilities, and absent cost-benefit analysis.

6. Cross-Protocol Consistency

Cross-protocol consistency addresses deception coherence across multiple services that sophisticated attackers probe systematically. The challenge manifests across four dimensions:

- identity consistency requiring hostname matching across SSH uname, MySQL variables, and SSL certificates
- version consistency demanding OS versions align logically (Ubuntu 18.04 shouldn't report Apache 2.4.52)
- filesystem consistency necessitating FTP-accessible files existing in SSH
- timing consistency requiring uniform geographic latency across protocols

Detection scenario: SSH reports Ubuntu 20.04, HTTP shows "nginx/1.14.0 (Ubuntu 18.04)", FTP returns "UNIX Type: L8", MySQL indicates Windows – these inconsistencies immediately identify honeypots to sophisticated attackers. Research findings: virtually zero academic attention, no systematic approaches, most honeypots deploy services independently. Practical challenges include managing consistency across 5+ protocols, the lack of centralized configuration systems, and the absence of validation tools. Assessment: high-priority gap for multi-protocol deployments as attackers increasingly employ multi-vector detection.

IV. EMPIRICAL VALIDATION STUDY

Our eight-day deployment captured 75,598 attacks, providing empirical validation and revealing real-world patterns informing deception requirements.

1. Attack Characterization and Protocol Distribution

Temporal analysis revealed remarkably aggressive attack onset: the first attack, targeting the exposed MySQL service, occurred precisely three minutes after honeypot exposure, demonstrating that internet-wide scanning discovers new targets nearly instantaneously. Within the first 24 hours, attackers generated 9,142 attacks, averaging 381 attempts per hour. Over the complete 195.35-hour deployment period, the cumulative total of 75,598 attacks averaged 387 per hour, or approximately 9,288 per day. This sustained high-intensity attack rate creates immediate implications for honeypot deception: no "warm-up" period exists during which configurations can be refined – systems must achieve full operational realism before exposure. The attack volume necessitates computationally efficient deception approaches, as resource-intensive techniques such as LLM inference would quickly become unsustainable under sustained load. Attack intensity patterns showed temporal variation with notable spikes. The initial 24-hour period exhibited a relatively uniform distribution of the 9,142 recorded attacks. The full eight-day timeline showed sustained baseline activity punctuated by a major spike on July 29, reaching approximately 8,000 attacks in a three-hour window, suggesting either a coordinated scanning

campaign or a surge in botnet activity. This pattern indicates honeypots encounter both continuous automated scanning and periodic intensive campaigns requiring deception capable of handling both steady-state and burst traffic. Geographic distribution analysis revealed attacker origins spanning at least ten countries with notable concentration patterns. The United States accounted for 24.65% of attacks, representing the largest single-source nation. China contributed 9.07%, consistent with documented large-scale scanning infrastructure. Nigeria, at 8.58%, reflects emerging attack infrastructure in developing nations. The Netherlands contributed 7.62%, likely representing commercial hosting infrastructure rather than Dutch attackers specifically. Other significant sources included Russia (6.48%), India (5.55%), South Africa (5.26%), Germany (4.59%), and Romania (3.83%), with the remaining 17.99% distributed across many additional nations. This geographic diversity creates important implications for timing deception: realistic latency simulation must account for round-trip times that scale with geographic distance (Section V-A). Protocol targeting distribution showed overwhelming SSH dominance: 72,668 attacks, representing 96.12% of total traffic, targeted the SSH service exclusively. MySQL received 1,466 attacks (1.94%), demonstrating sustained interest despite a lower percentage. DNS captured 1,114 queries (1.47%), primarily reconnaissance rather than exploitation attempts. FTP was targeted in 320 attacks (0.42%), reflecting the diminished relevance of legacy file transfer protocols. SMTP recorded merely 30 attempts (0.04%), suggesting mail relay abuse has become a comparatively rare attack vector. The SSH dominance validates the literature's extensive focus on SSH deception quality as the highest-return investment, given the concentration of attack volume. However, the nearly 2,930 non-SSH attacks remain nontrivial: they represent distinct attacker populations and techniques that require protocol-specific deception. Furthermore, evidence suggested cross-protocol probing, in which attackers successful in SSH compromise attempts subsequently explored MySQL and FTP services, reinforcing the cross-protocol consistency requirements identified in Section III-F.

2. SSH Attack Analysis

SSH attacks, enabled by complete session recording, revealed systematic patterns in username targeting, password strategies, and command execution. The 524 unique usernames attempted during deployment showed a clear preference for privileged accounts: root received 3,245 authentication attempts (24.21% of username-identified traffic), and admin received 1,566 attempts (11.68%), collectively representing 35.89% of privileged account targeting. Service-specific accounts, including Oracle (217 attempts), PostgreSQL (132 attempts), and Ubuntu (217 attempts), indicated that attackers were targeting database servers and specific Linux distributions. Generic accounts user (776), test (262), and guest (161) represented low-effort attempts at common credentials. The username "solana" appeared with 127 attempts, suggesting attackers are targeting cryptocurrency mining infrastructure. The substantial Other category, containing 6,556 attempts (48.91%), demonstrated the attacker's use of extensive username dictionaries beyond common patterns. These targeting patterns validate the interactive environment requirements in Section III-C: honeypots must populate realistic user mixes that combine privileged accounts, service accounts, and regular users in proportions matching simulated system types. However, an excessive number of privileged accounts is suspicious – real production systems typically limit root access and employ sudo mechanisms. Missing service accounts for corresponding services that ostensibly operate creates inconsistency: a system exposing PostgreSQL should contain the postgres user, and its absence, despite running the database service, is implausible. Password attack patterns revealed a mix of brute-force and credential-stuffing strategies. Simple numeric sequences dominated: 123456 appeared most frequently, followed by 123, 12345, 1234, 12345678, and 123456789. Dictionary words, including password, admin, ubuntu, and root, were highly frequent. Some sophisticated patterns emerged: FattMan1234567890 and lq4k@WSX suggested exploitation of a leaked credential database rather than pure brute-force generation. The string "solana" appeared in both the username and password lists, indicating the common username-equals-password attack pattern. The distribution across the 20 top passwords, while including thousands of unique attempts overall, demonstrates that attackers combine focused, common-credential attacks with broader dictionary coverage. Command execution analysis revealed systematic reconnaissance patterns consistent with automated exploitation frameworks. The most frequent command, uname -s -v -n -r -m, with 1,431 occurrences, gathered comprehensive system information, including kernel name, version, network name, release, and machine architecture, guiding

subsequent exploitation attempt selection based on identified vulnerabilities. Uptime calculation sequences appeared frequently: 132 instances of cat /proc/uptime | cut -d -f1 | cut -d. -f1 followed by AWK processing specifically designed to calculate system uptime in a human-readable format. These commands serve a dual purpose: legitimate system reconnaissance and honeypot detection through unrealistic uptime values or impossibly fast filesystem access.

Particularly revealing were explicit honeypot detection attempts: 110 instances of echo this_is_a_very_unique_test_string_4z > /tmp/honeypot_test_file.txt && cat /tmp/honeypot_test_file.txt && rm /tmp/ deliberately tested filesystem write persistence and read consistency. Honeypots with read-only or nonpersistent filesystems fail this test immediately. The command sequencing – write followed immediately by read – also tests response timing: real systems require microseconds to milliseconds for filesystem operations, including kernel VFS layer processing and potential disk I/O, while memory-only honeypot implementations might respond in nanoseconds, betraying an artificial implementation. This command analysis directly validates timing deception requirements from Section III-D. The /proc/uptime sequences demonstrate that attackers expect specific filesystem access patterns, including the existence of the/proc pseudo-filesystem and realistic content formatting. The write-read-delete sequences explicitly probe filesystem persistence. Complex pipelines that use multiple cut operations and AWK processing establish timing expectations: genuine systems incur process creation overhead, inter-process communication through pipes, and multiple context switches totaling tens of milliseconds, whereas honeypots might complete in single-digit milliseconds, revealing their artificial nature through measurable timing differences.

3. Secondary Protocol Findings

MySQL attacks (1,466) showed extreme privileged targeting: admin (51.46%), root (48.36%), collectively 99.82%. Commands revealed automated patterns: SET AUTOCOMMIT=0 (51 times), indicating SQL injection frameworks, show databases (8 times), show variables like '%version_compile_os%' (4 times), demonstrating explicit OS fingerprinting, creating a cross-protocol consistency requirement – MySQL-reported OS must align with SSH information. DNS queries (1,114) revealed 248 attempts (22.3%) targeting the version. bind – explicit version fingerprinting. Real servers frequently disable or return generic strings, so the guideline is: honeypot DNS must handle them realistically. Query types: ALL_RECORDS (390, 35.01%) performing reconnaissance, TXT (336, 30.16%), A records (208, 18.67%). FTP attacks (320) showed anonymous username dominance (30 attempts, 63.83%), www (15, 31.91%). Passwords followed anonymous conventions, with email formats accounting for 34.04%. SMTP attacks (30) showed tiscali.it domains dominated (13 instances, 43.3%), suggesting a specific campaign. Destination patterns: tiscali.it (13), hotmail.com (5), gmail.com (3), yahoo.com (2).

V. RESEARCH GAPS AND IMPLICATIONS

Synthesizing state-of-the-art with empirical patterns identifies critical gaps, prioritized by detection risk and research-maturity deficiencies. Figure 2 explicitly traces this synthesis, mapping the attack patterns observed in our deployment onto the deception dimensions they probe and the resulting priority gaps.

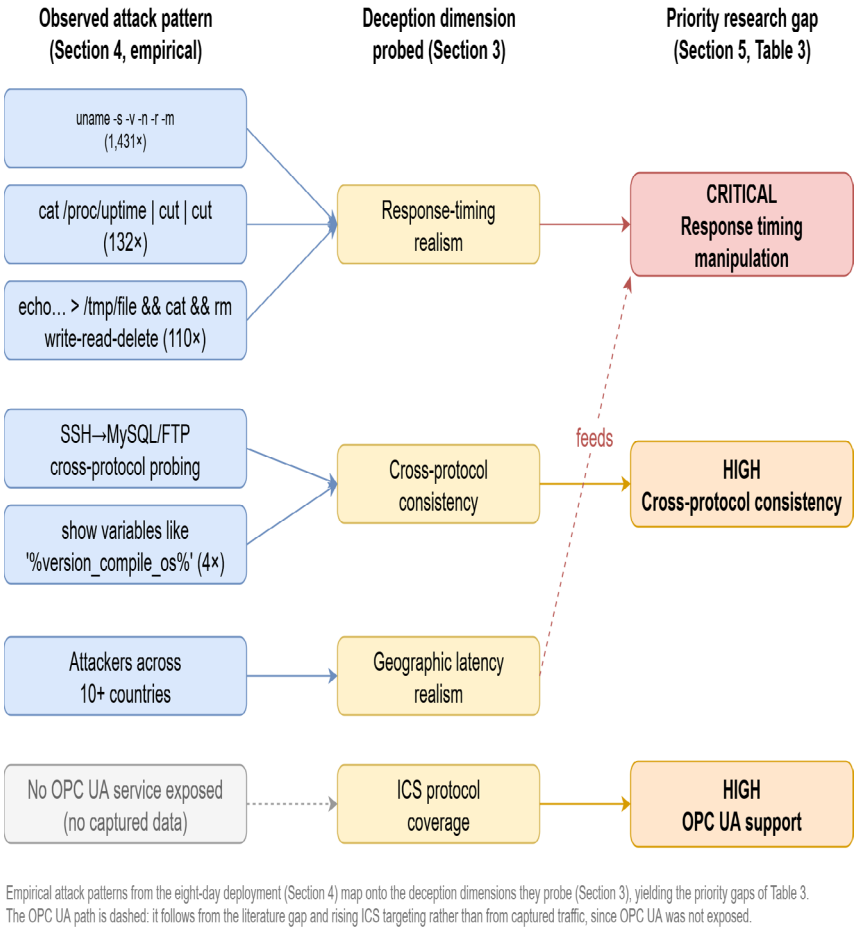


Figure 2. Empirical attack patterns from the eight-day deployment (Section IV) mapped onto the deception dimensions they probe (Section III) and the resulting priority research gaps (Table 3). The OPC UA path is dashed: it follows from the literature gap and rising ICS targeting rather than from captured traffic, since OPC UA was not exposed.

1. Critical Priority: Response Timing Manipulation

Response timing manipulation emerges as the highest-priority research gap through convergence of three critical factors: minimal existing research with only three significant works published over 20 years despite two decades of honeypot evolution (Section III-D), clear detection vector as evidenced by attacker command patterns specifically probing timing expectations through explicit filesystem tests and complex command sequences (Section IV-B), and universal applicability affecting all protocols and deployment contexts from traditional IT through ICS to emerging IoT environments. The empirical evidence from our deployment directly validates timing as a critical deception dimension that requires urgent research attention.

Attacker command sequences observed in our deployment reveal explicit timing expectations (Section IV-B). The cat /proc/uptime | cut | cut pipeline, executed 132 times, and the echo test > /tmp/file && cat /tmp/file write-test, executed 110 times, both impose timing expectations that honeypots routinely violate. On genuine Linux systems, the /proc pipeline completes in 15-50 milliseconds, depending on load and scheduling, while honeypots

answering from memory may complete it in <5 milliseconds. The write-read-delete sequence takes 10-100 milliseconds on real systems due to filesystem and disk overhead, yet memory-only honeypots may complete in under a millisecond. These differences form a measurable detection signal that timing analysis can exploit.

Our attackers spanned the USA (24.65%), China (9.07%), Nigeria (8.58%), the Netherlands (7.62%), and seven additional countries across four continents. This diversity creates a timing-consistency requirement that current honeypots cannot satisfy without explicit latency simulation. Real systems show network round-trip times proportional to geographic distance. European sources reaching European servers see <50ms, transatlantic US-Europe connections 100-150ms, and Asian-European paths 200-300ms. A honeypot that responds uniformly in <1ms regardless of source location exhibits physically impossible behavior, detectable through a simple latency measurement. Yet our literature review spanning 2004-2025 found zero implementations of geographic latency simulation, despite a straightforward approach: geolocate the attacker's IP and insert network-layer delays that match expected propagation delays.

Research opportunities span framework development, implementation techniques, and validation studies, creating a comprehensive research agenda. Framework development should produce a taxonomy categorizing timing deception approaches across multiple dimensions, including static delays with fixed response times independent of command characteristics, adaptive delays varying by command computational complexity to reflect genuine processing requirements, geographic-aware delays matching expected network latency based on attacker geolocation, and load-dependent delays reflecting multi-user contention and resource competition that real systems experience. Implementation research must develop timing profile libraries cataloging per-command response time distributions measured from real systems across diverse hardware configurations and load conditions enabling data-driven delay insertion, create transparent delay insertion mechanisms operating at network or application layers without requiring honeypot-specific modification to maintain compatibility across diverse implementations, and establish consistency maintenance techniques ensuring timing remains realistic across session duration avoiding suspicious timing variations between repeated commands that would betray artificial manipulation.

Validation studies require long-term deployments spanning weeks to months to measure whether timing deception reduces detection rates relative to nontiming-protected baselines. They also require attacker-behavior analysis to determine the tolerance thresholds separating acceptable delays from latencies that trigger disconnection, as well as adversarial testing of protected honeypots with timing-specific detection tools. Practical challenges remain. Threshold selection must balance sufficient delay for evasion against attacker patience, since excessive delay risks abandonment. Timing must stay consistent across repeated identical commands while still reflecting realistic load variation. Delay mechanisms add computational overhead to already-stressed high-interaction systems running full OS emulation. Finally, attacker patience varies: our data show both rapid initial attacks within 3 minutes of deployment, indicating impatient automated scanning, and sustained campaigns over the full 8-day period, indicating tolerance for moderate delays when a system appears genuine.

2. High Priority: Cross-Protocol Consistency and OPC UA Support

Cross-protocol consistency, identified in Section III-F as severely underexplored, gains practical validation through our multi-protocol deployment, which exposes five services simultaneously to internet traffic. The 2,930 non-SSH attacks (3.88%) and the SSH-to-MySQL/FTP probing reported in Section IV-A validate the consistency threat model: sophisticated attackers systematically probe multiple services, checking for inconsistencies that betray artificial environments through contradictory system characteristics that are impossible in genuine deployments.

The MySQL version fingerprinting command, show variables like '%version_compile_os%', observed four times in our captured traffic, creates a direct cross-protocol consistency requirement with SSH system information reporting. If the SSH uname command reports Ubuntu 20.04 while MySQL server variables indicate a Windows compilation environment, this immediate inconsistency exposes the honeypot implementation regardless of the individual protocol fidelity levels achieved through sophisticated emulation. Similarly, hostname consistency requirements demand that SSH hostname output generated by system calls matches MySQL @@hostname variable returned through database queries, DNS PTR records serving reverse IP lookup requests, and HTTP Host headers transmitted in web service communications. Our deployment using custom Cowrie configurations with modified filesystem and banner settings maintained basic consistency across exposed protocols. Still, many operational deployments using default settings across multiple independently configured protocol implementations likely exhibit detectable inconsistencies that sophisticated reconnaissance can identify.

Research opportunities include developing centralized identity and configuration management frameworks and ensuring that all exposed protocols reference a single authoritative source for hostname assignments, OS version reporting, network identity characteristics, and other shared system attributes that must remain consistent across protocol boundaries. Protocol interdependency modeling should systematically characterize which service combinations realistically coexist in real deployments: SSH plus a web server is an extremely common configuration present across millions of internet-exposed systems, whereas SSH plus multiple diverse ICS protocols might appear suspicious depending on the deployment context and organizational profile. Automated consistency validation tools that enable pre-deployment testing through systematic multi-protocol probing would prevent inconsistent deployments from reaching production exposure, where attackers can discover contradictions. The minimal existing research, combined with the clear practical necessity demonstrated by our empirical multi-protocol attack patterns, makes this a high-priority gap for any sophisticated honeypot deployment that exposes multiple services.

Our literature review found only minimal OPC UA coverage despite its strategic importance: few dedicated OPC UA honeypots exist [10],[11], even though OPC UA is now the modern ICS standard replacing legacy Modbus and S7comm across industrial deployments. Our deployment excluded OPC UA, so we collected no direct attack data, but the protocol's rapid industrial adoption and rising ICS targeting establish clear future relevance. OPC UA is technically demanding to emulate: the specification defines numerous service primitives, certificate-based authentication and encryption, and rich information modeling well beyond legacy protocols. The research opportunity is correspondingly strong. Only a handful of implementations exist, mature open-source libraries such as open62541 enable rapid prototyping, and the practical impact for critical infrastructure is high given OPC UA's growth across energy, manufacturing, and transportation. A credible approach mirrors the HoneyPLC S7comm strategy (Section III-B), which achieved deception fidelity against active fingerprinting and search engine indexing using the Snap7 library [9].

3. Medium Priority Gaps

Database protocol superficiality affects MySQL/PostgreSQL, which lack realistic query processing. Our 1,466 MySQL attacks employed simple probing, but increasing sophistication may demand deeper functionality, including query execution simulation, dynamic content generation, and transaction handling. Default configuration vulnerability represents a known issue with underutilized solutions (72% detectable [3]), indicating an operational gap requiring improved tooling to automate uniqueness. Temporal evolution addresses artifact aging, with limited research suggesting opportunities for frameworks. AI/LLM adversarial robustness concerns emerging approaches facing unexplored detection threats as attackers develop AI-specific techniques. Table 3 summarizes priorities.

Table 3. Research Gap Priority Matrix.

Gap Area	**Research Maturity**	**Detection Risk**	**Practical Impact**	**Priority**
Response Timing	Very Low	Very High	Universal	Critical
Cross-Protocol Consistency	Very Low	High	Multi-protocol	High
OPC UA Support	Low	Medium (Growing)	ICS Security	High
Database Protocol Depth	Low	Medium	Moderate	Medium
Default Config Avoidance	Known Issue	High	Implementation	Medium
AI Adversarial Robustness	Emerging	Future	AI Adoption	Medium

VI. DISCUSSION AND IMPLICATIONS

Our analysis reveals a paradox in honeypot deception research: the community has achieved remarkable sophistication in visible deception dimensions such as banner manipulation and protocol emulation, while neglecting invisible yet critical dimensions such as response timing, which sophisticated attackers increasingly exploit. This misalignment between research investment and attacker detection capability explains why 72% of deployments remain detectable despite two decades of technical advancement [3]. The detection arms race has shifted from "what honeypots say" to "how honeypots behave temporally and contextually." Our literature review and empirical validation show these dimensions remain severely underaddressed.

The convergence of our literature analysis with empirical attack patterns reveals several critical insights beyond individual gap identification. First, attacker sophistication demonstrates clear stratification: while 96.12% of attacks employ automated tools targeting SSH with predictable reconnaissance patterns, the remaining 3.88% exhibiting multi-protocol probing and timing-sensitive command sequences represent qualitatively different threats requiring fundamentally different defensive approaches. Current honeypot designs optimized against volume attacks remain vulnerable to quality attacks from sophisticated adversaries. Second, the operational-technical divide persists as a critical vulnerability: solutions exist for many detection problems (filesystem customization, artifact randomization), yet widespread adoption failures indicate that tooling and usability gaps matter as much as technical capability. Third, timing represents a universal detection vector cutting across all protocol domains and interaction types, yet has received essentially zero systematic research attention. This is not an incremental gap but a fundamental blind spot in the design of deception architecture.

The implications extend beyond honeypot deployments to broader deception system design. Any artificial system attempting to mimic genuine behavior faces timing consistency challenges: chatbots exhibiting inhuman response speeds, fake social media accounts posting with suspicious regularity, or simulated network traffic lacking realistic jitter all face analogous detection vulnerabilities. Our findings suggest a general principle: deception systems must address temporal characteristics with the same rigor as semantic content, as temporal patterns often provide simpler detection signals than content analysis.

Research priorities require strategic rebalancing. The critical timing gap demands immediate, coordinated effort, potentially through multi-institutional collaboration, given the scope spanning framework development, extensive real-system profiling, and long-term deployment validation. Cross-protocol consistency similarly benefits from standardization efforts, such as developing reference architectures and validation tools, thereby preventing ad-hoc, inconsistent implementations. In contrast, continued incremental improvements to mature areas like banner manipulation offer diminishing returns. The community should resist investing disproportionate effort in refining already-sophisticated capabilities while fundamental gaps persist.

The OPC UA gap exemplifies the broader challenge of keeping deception capabilities aligned with the evolving technology landscape. As industrial systems adopt modern protocols, legacy protocol honeypots become progressively less relevant regardless of implementation quality. This suggests the need for systematic horizon scanning to identify emerging protocols and technologies that require preemptive honeypot development before widespread targeting by attackers begins, rather than reactive development after exploitation campaigns emerge.

VII. CONCLUSION

This work provides a systematic understanding of honeypot deception capabilities and limitations through a dual-method analysis that combines a literature review with empirical validation, capturing 75,598 attacks. Our seven-category taxonomy reveals substantial variation in maturity, from highly developed banner manipulation to severely underexplored timing deception. The empirical deployment validates literature findings while revealing attack patterns that directly inform deception requirements: command sequences that explicitly test timing expectations and filesystem persistence demonstrate that attackers actively probe dimensions that current research neglects. The systematic gap analysis identifies response timing manipulation as the highest-priority research need, followed by cross-protocol consistency and OPC UA support. These gaps share a common characteristic: clear practical significance, as evidenced by our empirical data, combined with minimal existing research, creating a substantial opportunity for impactful contributions. The research roadmap spans immediate priorities (timing frameworks, consistency validation) and emerging opportunities (AI adversarial robustness). Organizations deploying honeypots face clear imperatives: prioritize SSH quality given the concentration of attacks, eliminate default configurations through systematic customization, and validate cross-protocol consistency in multi-service deployments. Researchers have a clear opportunity to address fundamental gaps with little competition and high practical impact. As attackers employ increasingly sophisticated multi-vector detection techniques, honeypot effectiveness depends on addressing not only visible deception dimensions but the invisible temporal and contextual characteristics that current approaches systematically neglect.

REFERENCES

[1]. N. Provos, “A virtual honeypot framework,” in Proc. 13th USENIX Security Symp., 2004, pp. 1–14.

[2]. P. Cao et al., “CAUDIT: Continuous auditing of SSH servers to mitigate brute-force attacks,” in Proc. 16th USENIX Symp. Networked Systems Design and Implementation (NSDI ’19), 2019. [Online]. Available: https://www.usenix.org/conference/nsdi19/presentation/cao

[3]. W. Cabral, C. Valli, L. Sikos, and S. Wakeling, “Review and analysis of Cowrie artifacts and their potential to be used deceptively,” in Proc. Int. Conf. Computational Science and Computational Intelligence (CSCI), 2019, pp. 166–171.

[4]. J. Franco, A. Aris, B. Canberk, and A. S. Uluagac, “A survey of honeypots and honeynets for Internet of Things, Industrial Internet of Things, and cyber-physical systems,” IEEE Commun. Surveys Tuts., vol. 23, no. 4, pp. 2351–2383, 2021, doi: 10.1109/COMST.2021.3106669.

[5]. Y. Zhou, “Chameleon: Towards an adaptive honeypot for the Internet of Things,” in Proc. ACM Turing Celebration Conf. China (ACM TURC ’19), 2019, pp. 1–5, doi: 10.1145/3321408.3321589.

[6]. M. Oosterhof, "Cowrie SSH/Telnet honeypot." Accessed: Nov. 1, 2025. [Online]. Available: https://github.com/cowrie/cowrie

[7]. "Conpot ICS/SCADA honeypot." Accessed: Nov. 1, 2025. [Online]. Available: https://github.com/mushorg/conpot

[8]. M. Dodson, A. R. Beresford, and M. Vingaard, "Using global honeypot networks to detect targeted ICS attacks," in Proc. 12th Int. Conf. Cyber Conflict (CyCon), 2020, pp. 275–291, doi: 10.23919/CyCon49761.2020.9131734.

[9]. E. López-Morales et al., "HoneyPLC: A next-generation honeypot for industrial control systems," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS '20), 2020, pp. 279–291, doi: 10.1145/3372297.3423356.

[10]. C.-A. Petre and A. Korodi, "Honeypot inside an OPC UA wrapper for water pumping stations," in Proc. 22nd Int. Conf. Control Systems and Computer Science (CSCS), 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8745269

[11]. O. Sassnick, G. Schäfer, T. Rosenstatter, and S. Huber, "A generative model-based honeypot for industrial OPC UA communication," in Computer Aided Systems Theory – EUROCAST 2024, 2024. [Online]. Available: https://arxiv.org/abs/2410.21574

[12]. "Endless: SSH tarpit." Accessed: Nov. 1, 2025. [Online]. Available: https://github.com/skeeto/endlessh

[13]. M. Sladić, V. Valeros, C. Catania, and S. Garcia, "LLM in the shell: Generative honeypots," in Proc. IEEE Eur. Symp. Security and Privacy Workshops (EuroS&PW), 2024, pp. 430–435. [Online]. Available: https://arxiv.org/abs/2309.00155

[14]. G. Wagener, R. State, A. Dulaunoy, and T. Engel, "Adaptive and self-configurable honeypots," in Proc. 12th IFIP/IEEE Int. Symp. Integrated Network Management (IM 2011), 2011, pp. 345–352.

[15]. S. Dowling, M. Schukat, and H. Melvin, "Using reinforcement learning to conceal honeypot functionality," in Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2018), ser. LNCS, vol. 11053. Springer, 2019, doi: 10.1007/978-3-030-10997-4_21.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Influence of Acquisition Trajectory on Area Target Recognition and Crime Scene Re-Localization Using Consumer LiDAR

Kagiso Lawrence Pitsoane
Department of Computer Science, University of the Western Cape
Cape Town, South Africa
6609194@myuwc.ac.za
ORCID:0009-0005-5516-4250

Mfundo Andrew Maneli
Department of Computer Science, University of the Western Cape
Cape Town, South Africa
amaneli@uwc.ac.za
ORCID:0000-0002-7080-5811

Omowunmi Elizabeth Isafiade
Department of Computer Science University of the Western Cape
Cape Town, South Africa
oisafiade@uwc.ac.za
ORCID:0000-0002-3028-6180

Abstract
Accurate crime scene documentation is essential for preserving evidentiary integrity and enabling reliable reconstruction. While 3D documentation technologies improve spatial representation, professional systems remain costly and operationally demanding—this limits adoption in resource-constrained environments such as South Africa. Consumer-grade Light Detection and Ranging (LiDAR)-enabled devices provide a potential alternative. However, the influence of acquisition strategies on forensic scene re-localization remains underexplored. This study investigates the effect of LiDAR acquisition trajectory on augmented reality (AR)-based crime scene re-localization. Vuforia Area Targets were generated from Apple iPad Pro devices. A controlled indoor mock crime scene was captured using three trajectories: horizontal sweep, circular orbit, and 45° oblique zig-zag. Experiments were performed on iPad Pro devices with M1, M2, and M4 processors across 45 independent trials. Performance was evaluated using Recognition Success Rate (RSR), Time-to-Localization (TTL), Tracking Stability (TS), and Evidence Recovery Rate (ERR). Results showed strong localization performance across all devices, with the iPad Pro M4 achieving 100% recognition success across all evaluated trajectories. The 45° oblique zig-zag trajectory produced the lowest mean localization time for all devices and achieved the highest or tied-highest ERR in two of the three devices. These findings provide preliminary evidence that consumer-grade LiDAR devices can support AR-based forensic scene re-localization and suggest that acquisition trajectory may influence downstream localization performance.

Keywords: Consumer-grade LiDAR, 3D crime scene reconstruction, AR-based localization, Acquisition trajectory, Digital forensics

I. INTRODUCTION

Accurate crime scene documentation is fundamental to preserving evidentiary integrity and enabling reliable scene reconstruction [1], [2]. Conventional methods, including digital photography, videography [3], manual sketching [4], and tape-based measurements [5], remain widely used, particularly in resource-constrained regions such as South Africa [6]. However, these approaches are susceptible to human error and often fail to capture the complete spatial context of a crime scene [7]. Three-dimensional (3D) crime documentation technologies, including terrestrial laser scanning (TLS), LiDAR, uncrewed aerial vehicles (UAVs), and photogrammetry, provide non-contact acquisition of spatial information and enable the creation of permanent digital crime scene records for reconstruction and forensic analysis [8], [9]. Despite their advantages, professional forensic scanners are expensive and require specialized expertise, limiting their adoption in resource-constrained regions [10]. Consumer LiDAR devices offer a portable and cost-effective alternative for digital crime scene documentation [11]. While previous studies have primarily focused on reconstruction accuracy [12-14], relatively little attention has been given to how acquisition strategies affect the practical usability of reconstructed scenes during forensic investigations. In practice, investigators must be able to

recognize and re-localize digitally documented scenes when revisiting evidence reliably. This study investigates the effect of acquisition trajectory on the recognition performance of Vuforia Area Targets generated using LiDAR-enabled iPad Pros within an indoor mock crime scene. Three scanning trajectories were evaluated: horizontal, circular, and 45° (üssü halka işleci° oblique. These configurations were evaluated based on scene recognition success (SRS), time to localization (TTL), and tracking stability (TS) under controlled object-removal conditions during re-visitation. The findings provide preliminary practical insights into optimizing LiDAR-based crime scene acquisition using consumer devices and demonstrate the potential of low-cost mobile LiDAR for forensic applications in resource-constrained environments. The remainder of this paper is organized as follows. Section II reviews the relevant literature and identifies the research gap this study addresses. Section III describes the research methodology, while Section IV presents and discusses the experimental results. Finally, Section V concludes the paper and outlines directions for future research.

II. RELATED WORK

This section reviews digital crime scene documentation, focusing on (i) the evolution from conventional documentation to 3D reconstruction, (ii) 3D acquisition technologies, and (iii) practical considerations affecting forensic deployment. These studies provide context for investigating acquisition trajectories using low-cost LiDAR devices. 3D crime scene documentation has evolved from conventional recording methods to spatial reconstructions that preserve scene geometry and spatial relationships for long-term analysis [7], [15]. Early systems employed integrated laser and structured-light scanning for virtual reality (VR) scene exploration [5], while photogrammetry improved accessibility through multi-spectral imaging and integration with post-mortem computed tomography [9], [16]. However, reviews indicate that no single acquisition method is universally optimal across forensic scenarios and operational constraints [17]. Primary technologies, including TLS, photogrammetry, micro-UAVs, and consumer mobile LiDAR, present trade-offs between accuracy, cost, and operational complexity [7], [8], [17–19]. TLS provides high geometric accuracy but is costly for routine field deployment [7], [18], whereas photogrammetry offers a flexible, cost-effective alternative but is sensitive to lighting and low-texture environments [9], [17]. Consumer mobile LiDAR enables rapid acquisition with sufficient geometric accuracy for many forensic applications, offering a practical balance of speed, cost, and usability [11], [19]. Recent work has also emphasized workflow integration and immersive scene re-visitation through systems such as the Forensic Holodeck and AR/VR frameworks [20], [21]. However, existing studies primarily evaluate acquisition methods in terms of geometric accuracy rather than downstream AR re-localization performance [17], [19]. From an AR and SLAM perspective, reliable re-localization depends on persistent visual correspondences, viewpoint coverage, and robust feature matching [22-25]. At the same time, place-recognition research shows that viewpoint, appearance, and environmental changes can affect the robustness of localization [26], [27]. How the acquisition trajectory influences subsequent AR re-localization, particularly in forensic scene documentation, remains insufficiently investigated. This study addresses this gap by comparing three LiDAR acquisition trajectories and their effects on Area Target recognition, localization time, tracking stability, and object identification.

III. METHODOLOGY

1. Experimental Environment

Experiments were conducted in a controlled indoor setting configured as a hypothetical crime scene. A 2 × 2 m evidence region, bounded by two adjacent walls, was configured with seven predefined evidence items, as shown in Fig. 1. The evidence layout remained unchanged during the Area Target acquisition and re-localization trials. A separate ERR evaluation subsequently used sequential object removal to assess evidence recovery. Each

acquisition trajectory was evaluated using identical experimental conditions on each device, with multiple independent trials performed to ensure repeatability.

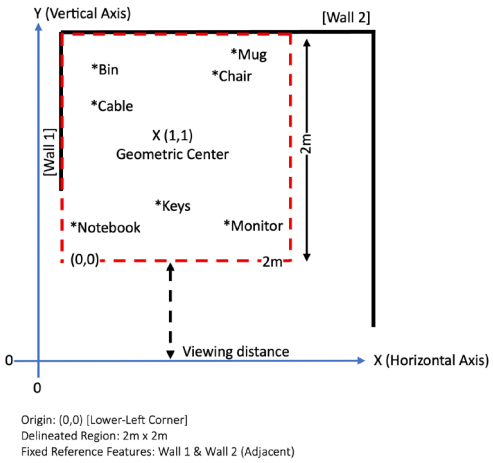


Figure 1. Top-down schematic of the $2\,\mathrm{m} \times 2\,\mathrm{m}$ experimental scene, showing the (0,0) coordinate origin, reference walls, and the distribution of evidence objects.

The lower-left corner was defined as the coordinate origin (0, 0), with the scene center located at (1, 1). Experiments were performed on Apple iPad Pro devices with M1, M2, and M4 processors, each equipped with an integrated LiDAR scanner. Their hardware specifications are summarized in Table 1.

Table 1. Camera and LiDAR specifications of the mobile devices used in this study.

Device	Rear Camera	LiDAR Scanner	Processor
iPad Pro 12.9-inch (M1)	12 MP Wide + 10 MP Ultra-Wide	Integrated LiDAR	Apple M1
iPad Pro 11-inch (M2)	12 MP Wide + 10 MP Ultra-Wide	Integrated LiDAR	Apple M2
iPad Pro 13-inch (M4)	12 MP Wide	Integrated LiDAR	Apple M4

For all trials, the device was held at approximately 1.5 m above ground and about 1 m from the scene center while maintaining a slow, constant walking speed.

2. Evaluation Metrics

Area Target performance was evaluated using four quantitative metrics over a fixed 30 s observation period (T_{observe}), recorded via Xcode on an Apple Mac mini (M2):

, sağ parantez ×100): Where büyük harf N alt s is the number of trials successfully recognized within 10 s and büyül trajectory condition.$= \frac{1}{N_s}\sum_{i=1}^{N_s} t_i$: The mean recognition time is for successful trials only. Trials exceeding 10 s were counted as unsuccessful for RSR and excluded from TTL.

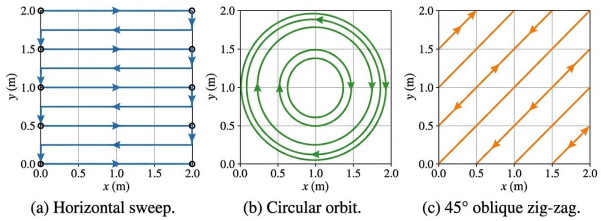
Eşittir $= \frac{\sum T_{\text{stable},i}}{N_s \times T_{\text{observe}}} \times 100$: Percentage of the 30 s observation period during which stable tracking was maintained, averaged across trials that successfully localized within 10 s.

, sağ parantez ×100): The percentage of the büyük harf E alt t predefined evidence items correctly identified (büyük in the localized AR scene.

Percentage of total predefined evidence items correctly identified in the localized AR scene. For each trial, the experimenter compared the localized AR scene against the predefined object inventory used as the ground truth. An object was counted as correctly recovered when its corresponding object was correctly identified in the localized AR scene. ERR was calculated as the percentage of correctly recovered objects divided by the total number of predefined objects. The same experimenter performed the experiments and the ERR scoring. No independent inter-rater assessment was conducted.

3. LiDAR Acquisition Trajectories

Three LiDAR acquisition trajectories were evaluated within the controlled crime scene to investigate the influence of capture geometry on Vuforia Area Target recognition and re-localization (Figure 2). The horizontal sweep provides uniform spatial coverage using five parallel passes, the circular orbit acquires multi-angle observations through concentric revolutions about the scene center, and the 4° oblique zig-zag maximizes angular viewpoint diversity through alternating diagonal traversals.



(a) Horizontal sweep. (b) Circular orbit. (c) 45° oblique zig-zag.

Figure 2. LiDAR acquisition trajectories were evaluated within the controlled 2 × 2 m mock crime scene. (a) Horizontal sweep provides uniform spatial coverage, (b) circular orbit maximizes multi-angle feature oobservation,and c) the 4, and (c) the 45° oblique zig-zag increases angular viewpoint diversity for robust Area Target generation and scene re-localization.

4. Experimental Procedure

The AR application was built in Unity 6 (v6000.0.78f1) using Vuforia Engine 11.4 and C# (Visual Studio 2021). Five independent trials were conducted per trajectory across the iPad Pro M1, M2, and M4 (45 total runs). Before each trial, the application was restarted, and the operator returned to the fixed starting position to maintain experimental independence.

IV. RESULTS AND DISCUSSION

This section discusses how consumer-grade LiDAR devices performed across various acquisition trajectories. As stipulated above, metric timings were recorded using Xcode on an Apple Mac mini (M2). Figure 3 depicts the controlled indoor setting configured as a hypothetical crime scene.

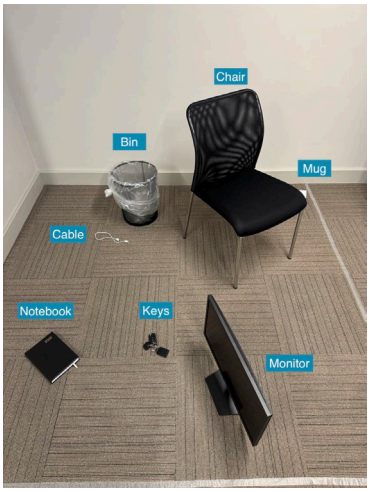


Figure 3. 2 m × 2 m experimental scene containing the distribution of evidence objects.

1. Area Target Recognition Performance

This subsection presents and discusses the results obtained from the RSR conducted on the acquired target areas for each device. Table 2 presents the recognition success rate and mean time-to-localization (TTL), with standard deviation (SD) calculated from successful trials.

Table 2. Recognition success rate and time-to-localization results.

Device	Acquisition Trajectory	Successful Trials / Total Trials	RSR (%)	TTL (s), Mean ± SD
iPad Pro M1	Horizontal Sweep	4/5	80	3.18 (±0.24)
	Circular Orbit	5/5	100	4.12 (±0.70)
	45° Oblique zig-zag	5/5	100	2.95 (±0.47)
iPad Pro M2	Horizontal Sweep	5/5	100	4.48 (±0.89)
	Circular Orbit	4/5	80	3.56 (±0.03)
	45° Oblique zig-zag	5/5	100	2.72 (±0.87)
iPad Pro M4	Horizontal Sweep	5/5	100	2.34 (±0.13)
	Circular Orbit	5/5	100	2.21 (±0.00)
	45° Oblique zig-zag	5/5	100	2.08 (±0.13)

Across all evaluations, the iPad Pro M4 achieved 100% RSR and the shortest TTL (2.08-2.34 s) among the tested devices and trajectory conditions. The 45° oblique zig-zag trajectory yielded the lowest mean TTL for each device in the present experiment (2.95 s for M1, 2.72 s for M2, and 2.08 s for M4), which may be associated with greater viewpoint diversity. However, viewpoint coverage and feature-level persistence were not directly measured.

Overall recognition was exceptional across all trials, with only two failures recorded: iPad Pro M1 dropped to 80% RSR on horizontal sweep, and iPad Pro M2 dropped to 80% RSR on circular orbit. All other device trajectory combinations achieved 100% RSR, shown in Figure 4.

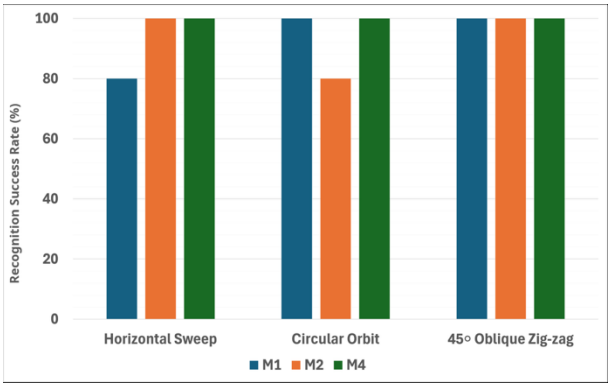


Figure 4. RSR comparison across all devices tested.

Generational comparison reveals a progressive improvement in localization efficiency. While the M1 and M2 devices exhibited greater variability in TTL across acquisition trajectories, the M4 maintained consistently low TTL values with minimal variation. The consistently low TTL values observed for the M4 indicate stable localization performance under the tested conditions; however, the contribution of hardware generation or processing capability cannot be isolated by the present experimental design.

2. Tracking Stability Analysis

This subsection presents the mean TS results for each device. Table 3 summarizes the mean TS values across all scans relative to the 30 s observation period and includes the corresponding percentages. Values closer to the full 30s observation period indicate greater TS and, consequently, better performance.

Table 3. TS results after successful localization of the area target.

Device	Acquisition Trajectory	Stable Tracking Time (s)	Observation Time (s)	TS (%)
iPad Pro M1	Horizontal Sweep	27.48	30	91.60
	Circular Orbit	26.74	30	89.13
	45° Oblique zig-zag	28.45	30	94.83
iPad Pro M2	Horizontal Sweep	27.89	30	92.97
	Circular Orbit	27.12	30	90.40
	45° Oblique zig-zag	29.45	30	98.17
iPad Pro M4	Horizontal Sweep	30	30	100
	Circular Orbit	30	30	100
	45° Oblique zig-zag	30	30	100

Figure 5 presents the same TS results expressed in seconds, providing a visual comparison across devices.

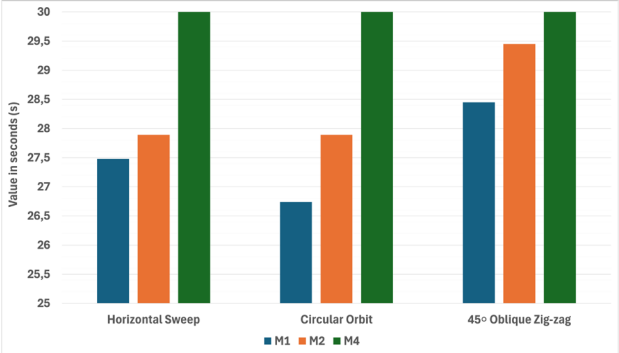


Figure 5. TS comparison across all devices used in testing.

TS results demonstrate consistently high performance across all devices and acquisition trajectories. Results also show that TS values ranged from 89.13% to 100%. The iPad Pro M4 achieved the highest stability, maintaining 100% TS across all trajectories. These results indicate stable tracking under the tested conditions, but field applicability in resource-constrained forensic environments requires further evaluation across varied scenes, operators, and environmental conditions. For the M1 and M2 devices, the 4° oblique zig-zag trajectory achieved the highest TS values of 94.83% and 98.17%, respectively, suggesting that diverse viewpoints enhance feature persistence and tracking reliability. Conversely, the circular orbit produced the lowest TS values for the M1 and M2 devices. Overall, the M4 showed more consistent tracking under the tested conditions. However, the effects of device generation, hardware differences, and processing capability could not be isolated in the present design.

3. Evidence Recovery Performance

This subsection presents the results of the ERR test, demonstrating the effectiveness of the proposed approach for supporting crime scene re-visitations in low-resource environments. In addition to quantifying the number of successfully recovered evidence items, the analysis considers instances of AR drift and other tracking anomalies observed during the experiments. Table 4 summarizes the detailed recovery results for each device and acquisition trajectory. Seven predefined scene objects were placed within the 2 × 2 m evidence region and sequentially removed during the ERR evaluation.

Table 4. ERR RESULTS AFTER SUCCESSFUL SCENE LOCALIZATION

Device	Acquisition Trajectory	Chair	Monitor	Keys	Note book	Cable	Mug	Bin	Recovered / Total	ERR (%)
iPad Pro M1	Horizontal Sweep	✓	✓	✓	✓	✓	✓	✓	7/7	100
	Circular Orbit	X	X	✓	✓	✓	X	✓	4/7	57.14
	45 ° Oblique zig-zag	✓	✓	X	✓	✓	✓	✓	6/7	85.71
iPad Pro M2	Horizontal Sweep	✓	✓	X	✓	X	X	✓	5/7	71.42
	Circular Orbit	✓	✓	✓	✓	X	X	✓	5/7	71.42

	45 ° Oblique zig-zag	✓	✓	✓	✓	✓	✓	✓	7/7	100
iPad Pro M4	Horizontal Sweep	✓	✓	✓	✓	✓	✓	✓	7/7	100
	Circular Orbit	✓	✓	✓	✓	X	✓	✓	6/7	85.71
	45 ° Oblique zig-zag	✓	✓	✓	✓	✓	✓	✓	7/7	100

Figure 6 provides a visual comparison of the corresponding ERR percentages across the evaluated devices and trajectories.

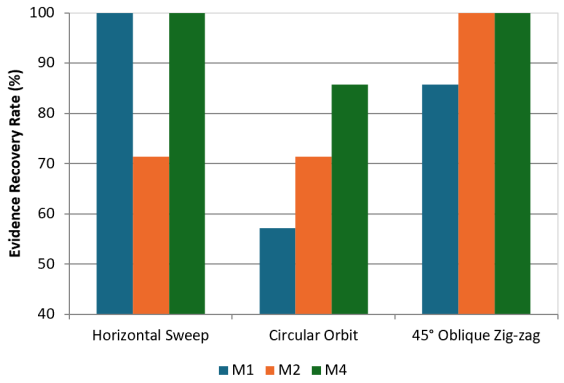
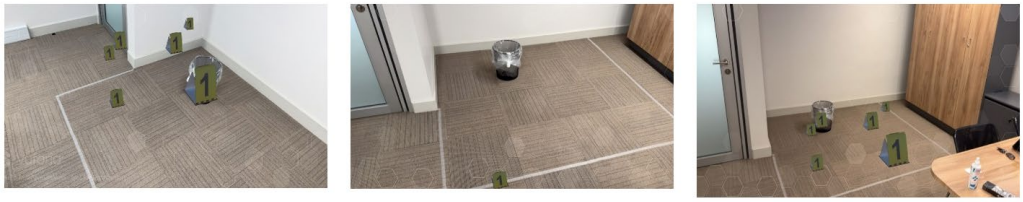


Figure 6. ERR comparison across all devices used in testing

Across the evaluated trajectories, the 45° oblique zigzag generally yielded the highest evidence-recovery performance. The iPad Pro M1 achieved 100% ERR for the horizontal sweep, 57.14% for the circular orbit, and 85.71% for the 45° oblique zig-zag. The iPad Pro M2 achieved 71.42% ERR for both the horizontal sweep and circular orbit, while the 45° oblique zig-zag achieved 100%. The iPad Pro M4 achieved 100% ERR for both the horizontal sweep and 45° oblique zig-zag, compared with 85.71% for the circular orbit. Overall, the 45° oblique zigzag showed the strongest evidence-recovery performance across the three devices. However, the results should be interpreted as trends under the tested conditions, given the limited number of trials per condition.

Figure 7 illustrates representative localization behaviors observed during ERR testing. Figure 7(a) shows the iPad Pro M1 after localization, where the virtual evidence markers remain visible but exhibit an orientation shift relative to the physical scene, indicating AR drift or axis misalignment. Figure 7(b) shows the iPad Pro M2 during a tracking-loss event, in which the virtual evidence markers are no longer rendered, indicating a temporary loss of AR tracking. Figure 7(c) shows the iPad Pro M4 under stable tracking, with the virtual evidence markers maintaining the expected spatial orientation relative to the physical scene. These observations provide qualitative context for the ERR results reported in Table 4.



(a) M1 iPad Pro (b) M2 iPad Pro (c) M4 iPad Pro

Figure 7. Representative AR localization behaviors observed during ERR testing: (a) orientation shift of virtual evidence markers on the iPad Pro M1, (b) temporary tracking loss on the iPad Pro M2, indicated by the disappearance of the virtual markers, and (c) stable tracking on the iPad Pro M4, with markers maintaining the expected spatial orientation.

These qualitative observations are consistent with the quantitative ERR results in Table 4, in which the M1 circular trajectory yielded the lowest ERR (57.14%). At the same time, the M2 and M4 achieved higher recovery under their respective evaluated trajectories. However, the observed anomalies were not independently quantified and should therefore be interpreted as qualitative observations rather than additional performance measures. The M4 results indicate stable tracking under the tested conditions, rather than general device superiority.

4. Impact of acquisition trajectory on LiDAR-based AR crime scene re-localization.

Beyond evaluating hardware performance, these findings demonstrate the viability of consumer-grade mobile LiDAR for digital crime scene documentation, where professional scanning systems are cost-prohibitive. The results indicate that the acquisition trajectory was associated with differences in observed localization and evidence recovery performance, with the 45° oblique zigzag and horizontal sweep trajectories generally yielding favorable results under the tested conditions. These findings suggest that the capture strategy may influence the reliability of AR-based scene re-localization under the tested conditions and motivate further evaluation in resource-constrained forensic settings.

5. Limitations

Several limitations should be considered. The study used one controlled indoor scene, one operator, one AR framework (Vuforia Engine), and five trials per device-trajectory condition, which limited generalizability and statistical inference. A single experimenter also scored ERR without independent inter-rater assessment. Future studies should include multiple scenes, operators, lighting conditions, AR frameworks, and larger trial sets to improve generalizability and statistical robustness.

V. SUMMARY AND CONCLUSION

This study investigated the influence of LiDAR acquisition trajectories on AR-based crime scene re-localization using iPad Pro devices with M1, M2, and M4 processors across 45 experimental trials. Results indicate that the acquisition trajectory was associated with differences in localization speed, tracking stability, and evidence recovery performance under the tested conditions. The 45° oblique zigzag trajectory produced the lowest mean TTL across all three devices (2.08-2.95 s), while the iPad Pro M4 achieved 100% RSR and TS across all evaluated conditions. These findings provide preliminary evidence of the feasibility of using consumer mobile

LiDAR for accessible digital crime scene documentation under the tested conditions, particularly in resource-constrained environments.

Future work will extend this evaluation to larger environments, variable illumination, occlusions, dynamic scene changes, and additional consumer LiDAR hardware. Additionally, future research will explore automated trajectory planning alongside cybersecurity frameworks, including encrypted transmission, authentication, and privacy-preserving data sharing, to support trusted, secure adoption in operational forensic processes.

REFERENCES

[1]. V. Berezowski, X. Mallett, and I. Moffat, "Geomatic techniques in forensic science: A review," Science & Justice, vol. 60, no. 2, pp. 99–107, 2020.

[2]. R. de Leeuwe, "The hiatus in crime scene documentation: Visualization of the location of evidence," Journal of Forensic Radiology and Imaging, vol. 8, pp. 13–16, 2017, special Issue: Papers from the joint ISFRI & IAFR Conference 2016.

[3]. D. Raneri, "Enhancing forensic investigation through the use of modern three-dimensional (3d) imaging technologies for crime scene reconstruction," Australian Journal of Forensic Sciences, vol. 50, no. 6, pp. 697–707, 2018.

[4]. T. Bhukya, "An Overview of Crime Scene Sketching," SSRN, 2023, doi: 10.2139/ssrn. 4461745.

[5]. J. Wang, Z. Li, W. Hu, Y. Shao, L. Wang, R. Wu, K. Ma, D. Zou, and Y. Chen, "Virtual reality and integrated crime scene scanning for immersive and heterogeneous crime scene reconstruction," Forensic Science International, vol. 303, p. 109943, 2019.

[6]. R. J. Mokwena, "Reconsidering the conventional reconstruction mechanism of a crime scene: Exploring forensic photography," *OIDA International Journal of Sustainable Development*, vol. 17, no. 12, pp. 99–108, 2024.

[7]. R. Tredinnick, S. Smith, and K. Ponto, "A cost-benefit analysis of 3D scanning technology for crime scene investigation," Forensic Science International: Reports, vol. 1, p. 100025, 2019, DOI: 10.1016/j.fsir.2019.100025

[8]. A. Sazaly, M. F. Mohd Ariff, and A. F. Razali, "3D Indoor Crime Scene Reconstruction from Micro UAV Photogrammetry Technique," Engineering, Technology & Applied Science Research, vol. 13, pp. 12 020–12 025, 12 2023, DOI:10.48084/etasr. 6260.

[9]. C. Villa and C. Jacobsen, The Application of Photogrammetry for Forensic 3D Recording of Crime Scenes, Evidence and People. Cham: Springer International Publishing, *Essentials of Autopsy Practice*, 8th ed., edited by Guy N. Rutty, 2019, pp. 1–18, DOI:10.1007/978-3-030-24330-2_1

[10]. C. Villa, N. Lynnerup, and C. Jacobsen, "A virtual, 3d multimodal approach to victim and crime scene reconstruction," Diagnostics, vol. 13, no. 17, 2023.

[11]. S. Kottner, M. J. Thali, and D. Gascho, "Using the iPhone's lidar technology to capture 3d forensic data at crime and crash scenes," Forensic Imaging, vol. 32, p. 200535, 2023, DOI:10.1016/j.fri.2023.200535

[12]. M. Lawas, S. Y. Williams, S. Jameson, A. R. Gonzalez, P. Ernst, and J. Donfack, "Assessing agreement among crime scene measurement methods," Journal of Forensic Sciences, vol. 67, no. 4, pp. 1715–1727, 2022.

[13]. M. Gregg, B. Muralikrishnan, and M. Shilling, "Toward a documentary standard for performance testing of terrestrial laser scanners used in forensic practice: A statistical procedure to assess change in instrument precision," Journal of Forensic Sciences, vol. 71, no. 2, pp. 1023–1031, 2026.

[14]. M. F. Emri, C. L. Lau, M. H. Razali, A. N. Idris, S. A. Sulaiman, K. M. Idris, and M. N. Abdul Ghani, "A comparative analysis of reconstructive crime scene documentation methods using lidar and surveying tools," Built Environment Journal, vol. 22, no. SI, pp. 37–53, 2025.

[15]. P. Ren, W. Shui, J. Liu, Y. Fan, W. Zhao, and M. Zhou, "A sketch-based 3D modeling method for 3d crime scene presentation," The Journal of Digital Forensics, Security and Law: JDFSL, vol. 13, no. 1, pp. 43–58, 2018.

[16]. G. Edelman and M. Aalders, "Photogrammetry using visible, infrared, hyperspectral and thermal imaging of crime scenes," Forensic Science International, vol. 292, pp. 181–189, 2018.

[17]. G. Galanakis, X. Zabulis, T. Evdaimon, S.-E. Fikenscher, S. Allertseder, T. Tsikrika, and S. Vrochidis, "A study of 3d digitisation modalities for crime scene investigation," Forensic Sciences, vol. 1, no. 2, pp. 56–85, 2021.

[18]. R. A. Azmil, M. F. Mohd Ariff, A. F. Razali, S. N. Azmy, N. Darwin, and K. M. Idris, "Transforming physical crime scene into geospatial-based point cloud data," Engineering, Technology & Applied Science Research, vol. 14, no. 3, pp. 13 974–13 981, Jun. 2024.

[19]. F. M. Sheshtar, W. M. Alhatlani, M. Moulden, and J. H. Kim, "Comparative analysis of lidar and photogrammetry for 3D crime scene reconstruction," Applied Sciences, vol. 15, no. 3, 2025.

[20]. L. C. Ebert, T. T. Nguyen, R. Breitbeck, M. Braun, M. J. Thali, and S. Ross, "The forensic holodeck: an immersive display for forensic crime scene reconstructions," Forensic Science, Medicine, and Pathology, vol. 10, no. 4, pp. 623–626, 2014.

[21]. I. Divya, B. Vybhavi, M. Marimuthu, B. Suchithra, R. Praveen, and N. Keerthana, "AI-Driven Forensic Investigation: Automated 3D Reconstruction and Evidence Visualization," in 2025 International Conference on Next Generation Information System Engineering (NGISE), vol. 1, March 2025, pp. 1–6, DOI:10.1109/NGISE64126.2025.11085283.

[22]. T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixé, "Understanding the limitations of CNN-based absolute camera pose regression," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3297–3307, doi: 10.1109/CVPR.2019.00342.

[23]. J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg, "Global localization from monocular SLAM on a mobile phone," *IEEE Trans. Visualization and Computer Graphics*, vol. 20, no. 4, pp. 531–539, 2014, doi: 10.1109/TVCG.2014.27.

[24]. R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015, doi: 10.1109/TRO.2015.2463671.

[25]. P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947, DOI:10.1109/CVPR42600.2020.00499.

[26]. R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307, DOI:10.1109/CVPR.2016.572.

[27]. A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 883–890.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

The Dependency of NIS2 Critical Sectors on Operational Technology: Implications for Cybersecurity Resilience and Critical Infrastructure Protection

Anes Mirojević
PhD Candidate, University of Buckingham
Sarajevo, Bosnia and Herzegovina
anes.mirojevic@stu.ssst.edu.ba
mirojevic.a@gmail.com
ORCID: 0009-0001-0587-225X

Jasmin Azemović, PhD
Professor, Sarajevo School of Science and Technology / University of Buckingham
Sarajevo, Bosnia and Herzegovina
Jasmin.azemovic@ssst.edu.ba
ORCID: 0009-0001-3319-5794

Abstract
Operational Technology (OT) plays a fundamental role in the operation of Critical Infrastructure (CI), yet its importance varies considerably across the sectors covered by the Network and Information Security Directive 2 (NIS2). This paper investigates the dependency of NIS2 Essential and Important sectors on OT systems and examines the implications for cybersecurity resilience and Critical Infrastructure protection. A literature-based Dependency of OT Assessment Framework (DOTAF) is proposed to evaluate OT dependency using four criteria: OT deployment, dependency on physical processes, operational impact of OT failure, and manual fallback capability. The framework is applied to all eighteen NIS2 sectors, demonstrating that thirteen exhibit medium, high, or very high dependency on OT. The findings further highlight the influence of ICT/OT convergence, cross-sector interdependencies, and real-world cyber incidents on operational resilience. The proposed framework provides a structured and reproducible approach for assessing OT dependency and supports risk-informed cybersecurity governance, resilience planning, and future sector-specific evaluations.

Keywords: Operational Technology, Critical Infrastructure, NIS2, Cybersecurity Governance, ICT/OT Convergence, Cyber Resilience

I. INTRODUCTION

The increasing digitalization of Critical Infrastructure (CI) has strengthened the integration of Information and Communication Technology (ICT) and Operational Technology (OT), improving operational efficiency while expanding the cybersecurity attack surface. CI, defined as the organizations, facilities, systems, and assets essential for the functioning of society and the economy [1], increasingly depends on OT systems to support the monitoring, control, and automation of physical processes [2]. As a result, cyber incidents affecting ICT environments can directly impact OT systems, creating consequences that extend beyond the digital domain and affect public safety, service continuity, and economic stability [3]. The Network and Information Security Directive 2 (NIS2) [4] establishes cybersecurity requirements for Essential and Important entities, many of which rely on OT to deliver critical services. Although international standards and the NIS2 Directive guide the securing of OT environments, they do not provide a structured methodology for assessing and comparing the level of OT dependency across the NIS2 sectors. This limits the ability to consistently evaluate sector-specific OT reliance and its implications for cybersecurity resilience. To address this research gap, this paper proposes the Dependency of OT Assessment Framework (DOTAF), a literature-based semi-quantitative framework for assessing and comparing the degree of OT dependency across all eighteen NIS2 sectors. The framework provides a structured, reproducible methodology for evaluating sector-specific OT dependencies and examining their implications for cybersecurity resilience and Critical Infrastructure protection. The contributions of this paper are threefold:



- It systematically analyses the role of OT within Critical Infrastructure sectors covered by NIS2.
- It proposes and applies the Dependency of OT Assessment Framework (DOTAF), a reproducible methodology for evaluating and comparing OT dependency across all eighteen NIS2 sectors.
- It demonstrates how varying levels of OT dependency influence cybersecurity resilience and Critical Infrastructure protection.

II. METHODOLOGY

The primary objective of this research is to examine the dependency of NIS2 Critical Infrastructure sectors on OT and assess the implications of varying levels of OT dependency for cybersecurity resilience. The study also examines internationally recognized standards, frameworks, and regulatory requirements relevant to OT security to support the development of a structured methodology for evaluating OT dependency across sectors. The research followed a four-stage methodology. First, a structured literature review was conducted to identify the characteristics of OT within Critical Infrastructure, using internationally recognized standards and guidance, including ISA/IEC 62443, NIST SP 800-82 Rev. 3, ENISA publications, and the NIS2 Directive. Second, common characteristics of OT dependencies were extracted from the reviewed literature. Third, these characteristics were translated into four evaluation criteria forming the Dependency of OT Assessment Framework (DOTAF). Finally, the framework was applied across all 18 NIS2 sectors to evaluate and compare their degrees of OT dependency. To support the development of DOTAF, the research reviewed internationally recognized OT cybersecurity standards, frameworks, and regulatory requirements, including ISA/IEC 62443, NIST SP 800-82 Rev.3, NIST CSF 2.0, and the NIS2 Directive [4], [5], [6], [7]. The study also examined ICT/OT convergence, representative OT cybersecurity incidents, and cross-sector dependencies to evaluate their implications for Critical Infrastructure resilience. Finally, existing governance frameworks, including ISO/IEC, ENISA, NIST, and NIS2 [4], [5], [6], [7], [8], were analyzed to support the development and application of the Dependency of OT Assessment Framework (DOTAF). This approach integrates technical, organizational, and regulatory perspectives to evaluate OT dependency and Critical Infrastructure resilience.

III. OPERATIONAL TECHNOLOGY AND CRITICAL INFRASTRUCTURE

1. Operational Technology Overview

OT refers to systems and devices used to automate, monitor, and control physical processes [2], [9]. OT systems interact directly with industrial equipment and infrastructure. They are commonly deployed in environments such as power generation facilities, manufacturing plants, water treatment facilities, transportation systems, and industrial processing plants [10]. The most prominent OT environments are Industrial Control Systems (ICS) [6], [9], which integrate hardware and software for monitoring and controlling industrial operations. Key components commonly found in OT environments typically consist of Programmable Logic Controllers (PLCs) [6], [11], Remote Terminal Units (RTUs) [12], Distributed Control Systems (DCS), Human-Machine Interfaces (HMI), Supervisory Control and Data Acquisition (SCADA) systems, Manufacturing Execution Systems (MES), Safety Instrumented Systems (SIS) [13]. These technologies collectively enable industrial organizations to automate processes, improve operational efficiency, and maintain reliable service delivery [6], [9]. Unlike ICT systems, OT systems are designed primarily to ensure operational continuity and the reliability of physical processes. Accordingly, protecting OT environments requires a different cybersecurity approach from that used in traditional ICT environments.

2. ICT and OT Convergence

Historically, ICT and OT [3] evolved as separate domains serving different operational purposes.

ICT systems primarily support data processing, information management, communication services, and business operations. OT systems primarily support industrial automation, physical process control, monitoring, and operational management. A fundamental difference between the two domains lies in their priorities. ICT environments prioritize confidentiality, integrity, and availability of information, while OT environments prioritize availability, reliability, and safety of physical processes. The principal differences between ICT and OT environments are summarized in Table 1.

Table 1. Top 5 Differences Between ICT and OT

No.	Aspect	ICT (Information and Communication Technology)	OT (Operational Technology)
1	Purpose and Function	Manages and processes digital data to support business operations, communication, and information flow.	Controls and monitors physical processes and machinery in industrial environments (e.g., power plants, pipelines, factories).
2	Core Priority	Focuses on the Confidentiality, Integrity, and Availability (CIA) of data and systems.	Prioritizes Availability, Safety, and Reliability of physical and industrial systems.
3	Cybersecurity Impact	Breaches cause financial loss, reputational damage, and data exposure.	Breaches can cause physical harm, operational shutdowns, or environmental hazards (e.g., radiation leaks, power outages).
4	Protection Objective	Aims to protect digital assets and sensitive data from unauthorized access or misuse.	Aims to protect physical assets and ensure operational continuity of industrial systems.
5	Operational Environment	Operates in corporate or enterprise ICT environments supporting users, databases, and networks.	Operates in industrial or critical infrastructure environments (e.g., power plants, water management facilities, manufacturing plants, utilities), managing real-world processes.

Despite these differences, the integration of ICT and OT [3] has accelerated over the last decade. Organizations increasingly connect OT environments to enterprise networks to improve visibility, efficiency, and decision-making capabilities.

Three forms of convergence can be identified as illustrated in Figure 1.

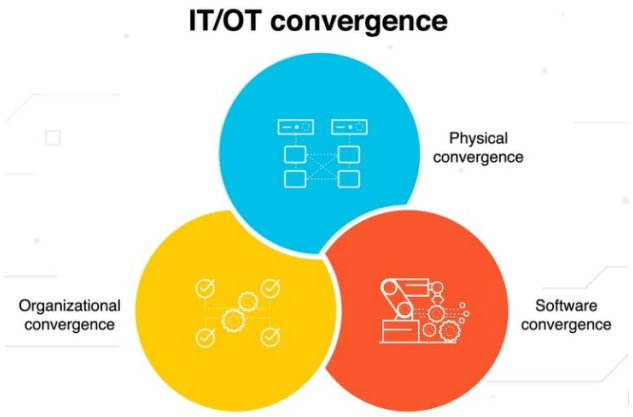


Figure 1. ICT/OT Convergence Types Required for Successful Operations [14]

The benefits of convergence include increased operational efficiency, enhanced visibility, predictive maintenance capabilities, and improved resource utilization. However, convergence also introduces cybersecurity challenges [3], [15] because vulnerabilities originating within ICT environments may affect OT systems and vice versa. Legacy OT systems, limited patching cycles, and increasing exposure of industrial networks further expand the attack surface.

IV. CYBERSECURITY GOVERNANCE FOR OPERATIONAL TECHNOLOGY

The protection of OT systems is supported by a combination of international standards, cybersecurity frameworks, and regulatory instruments [10]. The most prominent frameworks and standards for OT cybersecurity are summarized in Table 2.

Table 2. Principal Frameworks and Standards Supporting OT Cybersecurity

Framework / Standard	Organization	Latest Version	Primary Relevance to OT Security
ISA/IEC 62443 [5]	ISA / IEC	2024	International standard for securing Industrial Automation and Control Systems (IACS).
NIST SP 800-82 Rev.3 [6]	NIST	2023	Primary guidance for securing OT, ICS, SCADA, PLC, and DCS environments.
NIST CSF 2.0 [7]	NIST	2024	Cybersecurity governance and risk management framework.
ISO/IEC 27001 [8]	ISO / IEC	2022	Information security management standard applicable to OT environments.
ISO/IEC 27019 [16]	ISO / IEC	2024	Security guidance for energy-sector process control and SCADA systems.
ENISA Industry 4.0 [17]	ENISA	2019	EU guidance for cybersecurity in Industry 4.0 and industrial environments.

Framework / Standard	Organization	Latest Version	Primary Relevance to OT Security
MITRE ATT&CK for ICS [18]	MITRE	2025	Knowledge base of adversarial techniques targeting industrial control systems.
DOE C2M2 [19]	U.S. Department of Energy	2022	Cybersecurity capability maturity model for energy-sector organizations.
IEC 61511 / ISA 84	IEC / ISA	2025	A functional safety standard supporting the secure operation of industrial processes.
NIS2 Directive [4]	European Union	2022	Regulatory framework establishing cybersecurity obligations for Essential and Important entities.

Collectively, the ISA/IEC 62443 [5], NIST SP 800-82 [6], NIST CSF 2.0 [7], ISO/IEC 27001 [8], ENISA guidance [17], the NIS2 Directive [4], and recent Zero Trust guidance for Industrial IoT [20] establish complementary technical, operational, and governance requirements for securing OT environments and improving the resilience of CI.

The NIS2 [4] Directive introduces mandatory cybersecurity requirements for Critical Infrastructure entities operating within the European Union. The Directive establishes obligations related to:

- Risk management
- Incident reporting
- Governance and accountability
- Supervision and enforcement
- Cross-border cooperation

Together, these frameworks, complemented by emerging Zero Trust approaches for Industrial IoT [20], form a multi-layered approach to improving cybersecurity resilience across Critical Infrastructure sectors.

V. DEPENDENCY OF OT ASSESSMENT FRAMEWORK AND SECTOR EVALUATION (DOTAF)

A central finding of this research is that OT dependency varies significantly across sectors covered by NIS2 [4]. To provide a structured and repeatable assessment of OT dependency across NIS2 sectors, a semi-quantitative Dependency of OT Assessment Framework (DOTAF) was developed based on four criteria identified through the literature review, cybersecurity standards referenced in this paper, and information concerning the use of OT systems within CI sectors. DOTAF assesses the dependence of CI sectors on OT across four equally weighted dimensions: technological deployment, operational dependency, operational consequences, and resilience [21], [22].

The core baseline for the framework is based on the guidance contained in NIST Special Publication 800-82 Revision 3 – Guide to Operational Technology (OT) Security [6], the ISA/IEC 62443 series of standards [5], the

NIS2 Directive [4], and relevant ENISA publications [17]. The development process of the proposed DOTAF framework is illustrated in Figure 2.

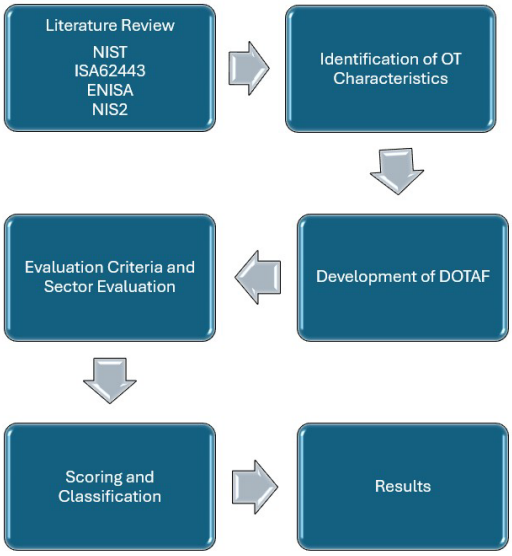


Figure 2. Development of DOTAF

The framework was developed based on the following evaluation criteria:

1. OT Deployment (A). This evaluates the extent to which a specific sector relies on PLCs, RTUs, DCS, SCADA, and other ICS technologies. (Score from 0-25 points)
2. Dependency on Physical Processes (B). This criterion assesses the extent to which the sector relies on automated physical processes controlled by OT systems. (Score from 0-25 points)
3. The Impact of OT failure on operations (C). This criterion evaluates the potential disruption caused by OT failure, including impacts on service delivery, public safety, the environment, and economic stability. (Score from 0-25 points)
4. Manual Fallback Capability (D). The ability of the organizations within a specific sector to continue operations through manual procedures in the event of OT systems failure. For Criterion D, higher scores indicate lower manual fallback capability and, therefore, greater dependence on OT. (Score from 0-25 points)

The scores are calculated using the following formula: Utilization of ICS systems (A) + Dependency on Physical Processes (B) + Impact of OT failure on operations (C) + Capability to fall back to manual operations (D). The maximum score possible is 100, while the minimum is 0. The four evaluation criteria were assigned equal weights of 25 points each because neither the reviewed cybersecurity standards nor the literature identifies any single criterion as consistently more influential than the others across all Critical Infrastructure sectors. Applying equal weighting, therefore, avoids introducing unsupported subjective bias while ensuring methodological transparency, consistency, and reproducibility of the assessment [21]. Criteria A, B, and C assess the technological and operational reliance of each sector on OT systems, with higher scores assigned to sectors that demonstrate greater dependence on OT for the delivery of essential services. Criterion D evaluates operational resilience by assessing a sector's practical ability to maintain operations through manual procedures following an OT system failure. Consequently, sectors with limited manual fallback capability receive higher scores.

The scoring methodology defines five levels of OT dependency based on the sum of the values assigned to each evaluation criterion:

- 81–100: Very High OT Dependency
- 61–80: High OT Dependency
- 41–60: Medium OT Dependency
- 21–40: Low-Medium OT Dependency
- 0–20: Low OT Dependency

Scoring methodology:

Each evaluation criterion is scored on a scale of 0 to 25, yielding a maximum possible OT Dependency Score of 100.

To reduce subjective judgment, the assigned score corresponds to predefined dependency levels derived from the documented operational characteristics of each sector [21].

The scoring criteria and corresponding decision rules applied consistently throughout the assessment are presented in Table 3.

Table 3. DOTAF Evaluation Criteria and Scoring Rules

Criterion	Score	Decision Rule
A. OT Deployment	0–5	Little or no documented use of OT. Operations are predominantly ICT-based.
	6–10	OT supports some operational activities but is not operationally significant.
	11–15	OT supports multiple operational functions but is not essential for core service delivery.
	16–20	OT is widely deployed across operational processes and is essential for most activities.
	21–25	Operations fundamentally depend upon extensive OT infrastructure.
B. Dependency on Physical Processes	0–5	Essential services do not depend on physical process control.
	6–10	Limited interaction with physical infrastructure.
	11–15	OT supports several automated physical processes.
	16–20	Most operational activities rely upon automated physical processes.
	21–25	Continuous control of physical processes is fundamental to sector operation.
C. Operational Impact of OT Failure	0–5	Minor operational disruption.
	6–10	Organizational disruption with limited external consequences.
	11–15	Significant regional operational disruption.
	16–20	National disruption affecting essential services.

Criterion	Score	Decision Rule
	21–25	Catastrophic consequences affecting public safety, environmental protection, economic stability, or national resilience.
D. Manual Fallback Capability	0–5	Operations can continue manually for extended periods.
	6–10	Manual operation remains feasible but with reduced efficiency.
	11–15	Only limited manual operation is possible.
	16–20	Manual operation can only temporarily maintain essential services.
	21–25	No practical manual operation is possible without OT.

The DOTAF scores and corresponding OT dependency levels for each NIS2 sector are presented in Table 4.

Table 4. Scores for Each NIS2 Sector Based on the Applied DOTAF Framework

Sector	A	B	C	D	Total	Level
Energy [6], [23]	25	25	25	20	**95**	Very High
Drinking Water [24]	24	24	24	20	**92**	Very High
Wastewater [24]	23	24	23	20	**90**	Very High
Manufacturing of Critical Products [17]	23	23	22	20	**88**	Very High
Chemicals [5]	22	23	22	18	**85**	Very High
Transport [25]	20	20	20	15	**75**	High
Waste Management	18	20	19	15	**72**	High
Space Infrastructure	17	18	20	13	**68**	High
Food Production [26]	16	18	18	13	**65**	High
Health	12	15	18	10	**55**	Medium
Digital Infrastructure [8]	10	14	18	10	**52**	Medium
Research Organizations	10	13	15	10	**48**	Medium
Postal and Courier Services	8	13	14	10	**45**	Medium
ICT Service Management [8]	6	10	12	7	**35**	Low–Medium
Banking	3	5	7	5	**20**	Low
Financial Market Infrastructure	2	5	6	5	**18**	Low
Public Administration	2	4	5	4	**15**	Low
Digital Providers [8]	1	3	4	4	**12**	Low

The sectors that have Very High OT dependency rely directly on SCADA, DCS, PLC, and other industrial control systems for their core operations.

Sectors classified as having **High OT Dependency** are essential for operational continuity and safety, although limited manual fallback mechanisms may exist.

Sectors with medium OT dependency are those where OT supports important operational functions but is not always the primary operational component.

Finally, sectors such as Banking, Financial Market Infrastructure, Public Administration, and Digital Providers exhibit relatively low OT dependency, as their core functions remain predominantly ICT-based.

The analysis indicates that thirteen of the eighteen sectors covered by NIS2 [4] demonstrate medium, high, or very high dependence on OT systems. This finding highlights the importance of OT security as a fundamental component of Critical Infrastructure protection.

The DOTAF framework is a literature-based evaluation framework rather than an empirical statistical model, and its analytical consistency was assessed by comparing it with internationally recognized OT guidance, sector-specific literature, and well-documented operational characteristics of CI sectors. The four evaluation criteria were compared against the principles described in NIST SP 800-82 Rev.3 [6], ISA/IEC 62443 [5], the NIS2 Directive [4], and ENISA publications [17]. The referenced standards and publications identify ICS deployment, dependency on physical processes, operational continuity, system availability, and resilience as key characteristics of OT environments. The four DOTAF criteria, along with their supporting standards and publications, are presented in Table 5.

Table 5. DOTAF Criterion References

DOTAF Criterion	Supporting reference
Extent of OT deployment	NIST SP 800-82 Rev.3, ISA/IEC 62443
Physical process dependency	NIST SP 800-82 Rev.3
Operational impact	NIST SP 800-82, ENISA, NIS2
Manual fallback capability	NIST SP 800-82, IEC 61511

The consistency between the DOTAF classifications and internationally recognized OT cybersecurity guidance supports the framework's suitability as a comparative assessment methodology for NIS2 sectors [22]. No significant inconsistencies were identified between the DOTAF classifications and the supporting references presented in Table 5. It is concluded that the DOTAF offers a reasonable and reproducible representation of OT dependency across NIS2 CI sectors.

VI. CROSS-SECTOR DEPENDENCIES AND INCIDENT ANALYSIS

1. Cross-Sector Dependencies

Critical Infrastructure sectors rarely operate independently. Many sectors depend directly or indirectly on services provided by other sectors. Energy systems depend on telecommunications, transportation systems depend on energy supplies, and healthcare organizations depend on both energy and communications infrastructure [27], [28].

These relationships create dependencies and interdependencies that amplify risk [27], [28], [29]. First-order dependencies represent direct service relationships between infrastructure providers and users. Second-order dependencies emerge when one infrastructure relies on another that, in turn, supports downstream services [27], [29]. As a result, disruptions affecting one sector may propagate across multiple sectors, increasing the societal

and economic impact of cyber incidents [27], [28], [29]. The interdependencies among Critical Infrastructure sectors and the potential propagation of disruptions across interconnected sectors are illustrated in Figure 3.

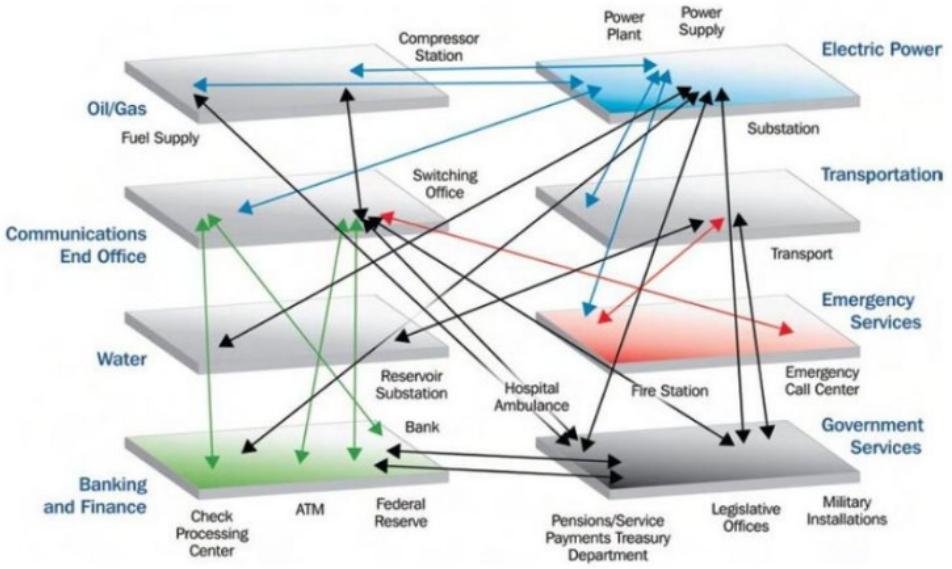


Figure 3. Interdependencies of Critical Infrastructure Sectors [29]

2. Incident Analysis

Several incidents illustrate the significance of OT security.

1) Stuxnet: The Stuxnet [6], [30] malware targeted industrial control systems associated with Iranian nuclear facilities. The malware manipulated PLC operations while simultaneously providing false operational feedback. This incident demonstrated that cyberattacks could directly affect physical industrial processes.

2) Ukrainian Power Grid Attacks: The attacks against the Ukrainian power grid [31], [32] in 2015 and 2016 demonstrated how attackers could exploit ICT systems to gain access to OT environments. The attacks disrupted electrical power delivery and affected hundreds of thousands of users.

3) Colonial Pipeline: The Colonial Pipeline [33] incident illustrated how attacks targeting ICT systems can indirectly affect OT operations. Although the ransomware attack targeted business systems, operational shutdowns were implemented as a precaution, resulting in fuel supply disruptions.

4) Bremanger Dam Incident: The compromise of systems associated with the Bremanger Dam [34] demonstrated the potential consequences of unauthorized access to industrial control systems. The incident highlighted the relationship between OT security and public safety.

Collectively, these incidents demonstrate that cyberattacks targeting OT environments may have consequences that extend beyond information systems to public safety, environmental protection, and economic stability.

VII. FINDINGS AND DISCUSSION

The analysis produced five primary findings.

Finding 1: Most NIS2 [4] sectors depend on OT. The majority of sectors covered by NIS2 [4] rely on OT systems for either core operations or essential supporting functions.

Finding 2: ICT/OT convergence expands the cybersecurity attack surface. The integration of ICT and OT systems increases operational efficiency but simultaneously introduces additional pathways through which cyber threats may affect industrial environments [3], [15].

Finding 3: OT disruptions create physical consequences. Unlike traditional ICT incidents, attacks on OT systems may disrupt physical processes and have consequences for public safety, environmental protection, and economic stability.

Finding 4: Cross-sector dependencies amplify cyber risk. Dependencies among Critical Infrastructure sectors increase the potential for cascading effects when disruptions occur [27], [28], [29].

Finding 5: Effective Critical Infrastructure protection requires OT-focused governance. The findings indicate that cybersecurity governance frameworks must address OT-specific requirements and support cross-sector collaboration.

Taken together, these findings demonstrate that OT dependency is not uniform across NIS2 sectors and that cybersecurity governance should therefore reflect differences in sector-specific operational reliance. Sectors with high OT dependency require particular attention to availability, safety, resilience, and continuity of physical processes, while lower-dependency sectors remain predominantly influenced by ICT-related risks.

VIII. LIMITATIONS

The proposed DOTAF framework is based on publicly available literature, international standards, and documented operational characteristics rather than empirical operational data obtained directly from CI operators. Consequently, the assessment reflects sector-level OT dependency and should not be interpreted as an organization-specific maturity assessment. Future work should validate the framework through expert surveys, industry collaboration, and quantitative weighting techniques.

IX. CONCLUSION

This paper examined the role of OT in protecting Critical Infrastructure sectors covered by the NIS2 [4] Directive. The analysis demonstrated that OT systems form the operational foundation of many sectors containing entities classified as Essential or Important under NIS2. The convergence of ICT and OT [3] environments provides substantial operational benefits but also increases cybersecurity risk. Real-world incidents demonstrate that disruptions to OT systems can have consequences that extend beyond cyberspace and directly affect society, public safety, and economic stability. The findings further indicate that most NIS2 sectors rely significantly on OT systems and that effective cybersecurity governance must therefore incorporate OT-specific security measures, resilience planning, cross-sector collaboration, and modern security architectures such as Zero Trust [20]. The developed DOTAF framework provides a systematic and reproducible methodology for evaluating OT dependency levels across Critical Infrastructure sectors covered by NIS2. It may serve as a foundation for future comparative, sector-specific risk assessments. Future efforts should continue to strengthen governance frameworks, promote information sharing, and improve the resilience of OT environments that support Critical Infrastructure.

REFERENCES

[1]. IBM, "Critical Infrastructure," 2025. [Online]. Available: https://www.ibm.com/think/topics/critical-infrastructure. [Accessed: Mar. 15, 2026].
[2]. National Institute of Standards and Technology, "Operational Technology," *NIST Computer Security Resource Center Glossary*, 2025. [Online]. Available: https://csrc.nist.gov/glossary/term/operational_technology. [Accessed: Apr. 25, 2026].
[3]. G. Murray and M. N. Ngo, "The convergence of IT and OT in critical infrastructure," in *Proc. Australian Information Security Management Conference*, Perth, Australia, 2017.
[4]. European Parliament and Council of the European Union, "Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)," *Official Journal of the European Union*, Dec. 2022.
[5]. International Society of Automation, "ISA/IEC 62443 Series of Standards," Oct. 2025. [Online]. Available: https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards. [Accessed: Jun. 15, 2026].
[6]. K. Stouffer, M. Pillitteri, S. Lightman, M. Abrams, and A. Hahn, *Guide to Operational Technology (OT) Security*, NIST SP 800-82 Rev. 3. Gaithersburg, MD, USA: National Institute of Standards and Technology, Sep. 2023.
[7]. National Institute of Standards and Technology, *Cybersecurity Framework (CSF) 2.0*. Gaithersburg, MD, USA: NIST, Feb. 2024.
[8]. International Organization for Standardization, *ISO/IEC 27001:2022 Information Security, Cybersecurity and Privacy Protection — Information Security Management Systems — Requirements*. Geneva, Switzerland: ISO, 2022.
[9]. M. Cook, A. K. Marnerides, C. Johnson, and D. Pezaros, "A Survey on Industrial Control System Digital Forensics: Challenges, Advances and Future Directions," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 3, pp. 1705–1747, Aug. 2023, doi: 10.1109/COMST.2023.3264680.
[10]. M. M. Aslam, A. Tufail, R. A. A. H. M. Apong, L. C. De Silva, and M. T. Raza, "Scrutinizing Security in Industrial Control Systems: An Architectural Vulnerabilities and Communication Network Perspective," *IEEE Access*, vol. 12, 2024.
[11]. ISA Global Cybersecurity Alliance, *Industrial Automation and Control System Taxonomy*. Research Triangle Park, NC, USA: ISA, 2021.
[12]. T. Goncalves, "Remote Terminal Units (RTU)," Aug. 15, 2023. [Online]. Available: https://fiixsoftware.com/glossary/what-are-remote-terminal-units/. [Accessed: Mar. 15, 2026].
[13]. W. Johnson and R. Roberts, "Understanding ISA-84 and Functional Safety in the Process Industries," 2025. [Online]. Available: https://blog.isa.org/understanding-isa-84-functional-safety-process-industries-manufacturing. [Accessed: Jun. 15, 2026].
[14]. Palo Alto Networks, "What Is IT/OT Convergence?" 2025. [Online]. Available: https://www.paloaltonetworks.com/cyberpedia/what-is-it-ot-convergence. [Accessed: Apr. 25, 2026].
[15]. KPMG, *Industrial Control System (ICS) or Operational Technology (OT) Threat Landscape*. New York, NY, USA: KPMG Assurance and Consulting Services LLP, Apr. 2024.
[16]. ISO/IEC 27019:2024, Information Security, Cybersecurity and Privacy Protection — Information Security Controls for the Energy Utility Industry. Geneva, Switzerland: International Organization for Standardization, 2024.
[17]. ENISA, Industry 4.0 Cybersecurity: Challenges & Recommendations, May 2019.
[18]. MITRE Corporation, "ATT&CK for ICS Matrix," 2025. [Online]. Available: https://attack.mitre.org/matrices/ics/. [Accessed: Jun. 15, 2026].
[19]. U.S. Department of Energy, *Cybersecurity Capability Maturity Model (C2M2)*, 2025. [Online]. Available: https://www.energy.gov/ceser/cybersecurity-capability-maturity-model-c2m2. [Accessed: Jun. 15, 2026].

[20]. "Blending Shared Responsibility and Zero Trust to Secure the Industrial Internet of Things," *IEEE Security & Privacy*, vol. 22, no. 5, pp. 96–102, Sept.–Oct. 2024.
[21]. A. O. Rotibi, N. Saxena, P. Burnap, and A. Tarter, "Extended Dependency Modeling Technique for Cyber Risk Identification in ICS," *IEEE Access*, vol. 11, pp. 37229–37242, 2023, doi: 10.1109/ACCESS.2023.3263671.
[22]. A. O. Rotibi, N. Saxena, P. Burnap, and C. Read, "System-Level Operational Cyber Risks Identification in Industrial Control Systems," *Cyber-Physical Systems*, vol. 11, no. 3, pp. 265–296, 2025.
[23]. Y. Wang, "Smart grid, automation, and SCADA systems security," *World Scientific Review*, vol. 9, Oct. 2012.
[24]. R. Teixeira and O. Costa, *Water Security Plan: Implementation Manual for Drinking Water Systems*. Luxembourg: Publications Office of the European Union, 2022.
[25]. L. Cebulski and S. Bernard, *Taking Cybersecurity Challenges into Account in Railway Safety*. Amiens, France: Établissement Public de Sécurité Ferroviaire, 2022.
[26]. D. Bertolozzi-Caredio et al., *Risks and Vulnerabilities in the EU Food Supply Chain*. Luxembourg: Publications Office of the European Union, 2023.
[27]. L. P. Lewis and F. Prochaska, *Critical Infrastructure Interdependency Analysis: Operationalizing Resilience Strategies*. Argonne, IL, USA: Argonne National Laboratory, U.S. Department of Energy, 2018.
[28]. T. R. Sonesson and J. Johansson, "Governance and interdependencies of critical infrastructures: Exploring mechanisms for cross-sector resilience," *Safety Science*, vol. 142, Art. no. 105383, 2021.
[29]. M. Ehlen and V. Novoselac, "Multi-hazard, multi-infrastructure, economic scenario analysis," *Environment Systems and Decisions*, vol. 33, no. 3, pp. 297–309, 2013.
[30]. Kaspersky, "What is Stuxnet?" 2025. [Online]. Available: https://www.kaspersky.com/resource-center/definitions/what-is-stuxnet. [Accessed: Mar. 15, 2026].
[31]. J. S. Beach-Westmoreland, *When the Lights Went Out: A Comprehensive Review of the 2015 Attacks on Ukrainian Critical Infrastructure*. McLean, VA, USA: Booz Allen Hamilton, 2019.
[32]. J. Slowik, *CRASHOVERRIDE: Reassessing the 2016 Ukraine Electric Power Event as a Protection-Focused Attack*. Hanover, MD, USA: Dragos Inc., 2019.
[33]. M.-A. Russon, "Colonial Pipeline cyberattack explained," BBC News, May 10, 2021. [Online]. Available: https://www.bbc.com/news/business-57050690. [Accessed: Apr. 25, 2026].
[34]. N. Adomaitis, "Norway spy chief blames Russian hackers for dam sabotage in April," Reuters, Aug. 13, 2025. [Online]. Available: https://www.reuters.com/technology/norway-spy-chief-blames-russian-hackers-dam-sabotage-april-2025-08-13/. [Accessed: Apr. 25, 2026].

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Learnscrolling: Harnessing Attention Economy Mechanics for Incidental Learning via Synthetic Social Media

Edward Abel
University of Southern Denmark
Denmark
abel@sdu.dk
ORCID: 0000-0002-3694-5116

Abstract
The average person spends over two hours daily on social media, much of it in passive, low-value doomscrolling that users report as unsatisfying yet find difficult to stop. These feeds are engineered for engagement through mechanisms including algorithmic personalization, social proof, variable reward schedules, and frictionless content delivery. This paper explores repurposing such mechanisms for learning within a synthetic social media platform, a feed where every participant is a large language model agent, tuned to a single real user who is the sole genuine node. While such infrastructure could manufacture an unhelpful echo chamber of synthetic consensus, our focus is on its potential to deliver incidental education through what feels like natural social scrolling. We present a conceptual framework and proof-of-concept prototype grounded in established educational principles including social learning theory, spaced repetition, and stealth education. Where existing agent simulations typically study collective dynamics among agents, we focus on a single user as the sole human on a personalized learning platform. The framework and prototype demonstrate how agents with distinct personas can introduce, scaffold, and reinforce learning objectives through varied feed interactions, suggesting a mechanism for incidental learning utilizing the same mechanics that currently drive doomscrolling.

Keywords: synthetic social media, LLM agents, filter bubbles, incidental learning, stealth education, attention economy, dual-use AI

I. INTRODUCTION

Social media users now average over two hours of daily platform use [1], yet a growing body of evidence suggests much of this time is spent in compulsive, low-value scrolling, a phenomenon Sharma et al. [2] term doomscrolling, that users themselves recognize as unrewarding but struggle to resist. This paradox is not accidental. The feeds driving this behavior are deliberately engineered for engagement rather than user satisfaction [3], deploying algorithmic personalization [4], variable reward schedules [5], social proof signals, and frictionless infinite scrolling that eliminates natural stopping points. The result is an infrastructure optimized to capture human attention, regardless of the value it delivers [6]. This paper asks: what if these same mechanisms were repurposed for learning. We explore this question through synthetic social media, a platform where every peer, commenter, and voice is a Large Language Model (LLM) agent, tuned to a single real user who is the sole genuine node (Figure 1). The technical components for such a platform already exist and are cheap to assemble [7], [8]. Our proposal configures these agents for learning: they teach, scaffold, and model reasoning through what feels like natural social interaction, delivering incidental education [9] via the same feed mechanics that drive doomscrolling, drawing on established principles of social learning, spaced repetition, and stealth education. We acknowledge that the same infrastructure could equally be configured for harm, manufacturing synthetic consensus rather than teaching, a concern that parallels the contested AI companion debate [10], [11], [12]. Our focus, however, is on the educational application. The paper presents a conceptual framework for learning-oriented synthetic social media that bridges the currently separate literatures on the

attention economy, AI companions, and educational technology. We examine how such a platform operates across epistemic, psychological, and social dimensions, and demonstrate feasibility through a proof-of-concept prototype showing how LLM agents can deliver incidental learning through familiar feed interactions.

II. BACKGROUND AND RELATED WORK

This section situates learning-oriented synthetic social media at the intersection of five areas: the attention economy (Section II.1), filter bubbles (Section II.2), AI companions (Section II.3), LLM agent simulations (Section II.4), and educational foundations (Section II.5). Section II.6 identifies the gap.

1. The Attention Economy and Doomscrolling

Having established the scale and persistence of low-value social media consumption, we turn to the mechanisms that sustain it. Zuboff [3] characterizes the underlying business model as surveillance capitalism, in which user attention is the primary commodity extracted and sold. Fogg [5] identified how computers function as persuasive technologies, deploying triggers and reinforcement to shape behavior. Wu [6] traces the broader history of the attention economy, arguing that commercial capture of human attention has intensified with each new medium. That the daily average of approximately two hours and twenty minutes [1] has remained stable despite growing public awareness [2] underscores the effectiveness of these mechanisms.

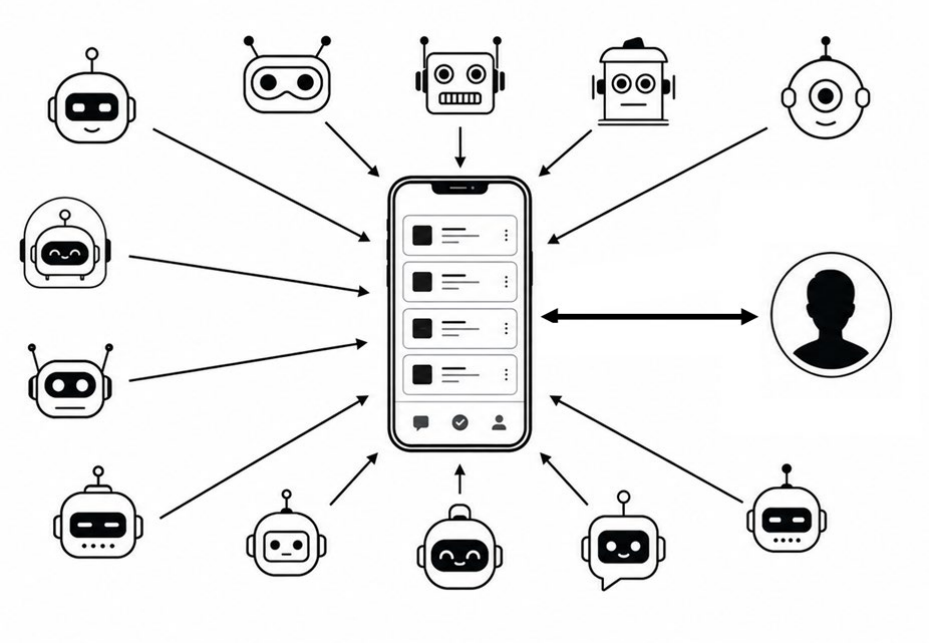


Figure 1. Illustration of the notion of a synthetic social media platform made up of many AI agents and a single human user

Critically, these mechanisms are format-neutral: what captures attention toward low-value content could capture it toward educational content instead, provided the learning is embedded in the experience rather than imposed on it.

2. Filter Bubbles and Echo Chambers

Pariser [4] introduced the filter bubble to describe how algorithmic curation creates personalized but increasingly isolated online experiences. Nguyen [13] distinguished between epistemic bubbles, where relevant voices are simply absent, and echo chambers, where outside sources are actively discredited. Chitra and Musco [14] formalized the relationship between recommendation-driven filtering and polarization, while Bruns [15] argued that empirical evidence for strong filter bubble effects is more contested than popularly assumed. All this work, however, shares a foundational assumption: the network is populated by real people. Synthetic social media removes the real voices entirely, the bubble is no longer a filter applied to reality but a replacement for it; one that could be designed to reinforce or to teach.

3. AI Companion Applications

AI companion applications have generated genuinely mixed evidence: heavy use has been linked to greater loneliness and dependency, lower well-being, and maximal attachment without psychosocial benefit [11], [12], [16], and chatbot engagement has been documented contributing to delusional thinking [17]. Yet loneliness carries serious mortality risks [10], and the theoretical roots of parasocial relationships are well established [18], [19]. The evidence is dose-dependent, moderate use may help, heavy use more likely harms. For this paper, the key observation is structural: this literature is entirely dyadic—one user, one bot—and does not examine what happens when the same logic scales to an entire platform, nor does it consider educational applications.

4. LLM Agent Social Simulations

A rapidly growing body of work uses LLM agents to simulate social media dynamics. Agent-only simulations constitute the majority, ranging from reusable platform digital twins [7] to large-scale simulations of 10,000+ agents [20] and comparative studies of recommendation algorithms and echo-chamber formation [21], [22], [23]. In all cases the research question concerns emergent dynamics among agents with no real user embedded. Human-agent studies are rarer. Donkers and Ziegler [24] placed 122 participants in a synthetic network populated by LLM agents, but their design uses multiple humans, uniformly polarized agents, no adaptation loop, and an experimental framing. Adversarial studies examine agents as threats, influence bots [25], and coordinated propaganda [26]. Across these categories, existing work does not examine a single user as a sole human on a personalized platform for educational applications.

5. Incidental Learning and Stealth Education

The learning platform we propose draws on established educational traditions, distinguishing three constructs: incidental learning, the acquisition of knowledge as a byproduct of another activity, without deliberate intent [9]; stealth learning, the deliberate embedding of instructional content within a non-instructional format [27]; and stealth assessment, the invisible measurement of performance within that activity [28]. These map respectively to the user's overall experience, the content-delivery design, and the Stage 3 measurement mechanism. Incidental learning is well-documented in media contexts [9], and social media specifically blurs formal and informal learning boundaries [29], [30]. Bandura's social learning theory [31] provides the core mechanism: people learn by observing others' model behavior. In a synthetic feed, agents serve as social models demonstrating and debating concepts. Bandura's work on self-efficacy [32] adds that vicarious experience builds learner confidence. The spacing effect maps directly onto the feed format, Cepeda et al. [33] found spaced practice outperformed massed practice in the vast majority of cases, and a feed naturally resurfaces concepts at intervals. Stealth learning, educational content disguised within entertainment or interactive contexts, has been

explored extensively in games. Shute and Ventura [28] developed the foundational framework for embedding invisible assessment within gameplay, while Sharp [27] names the phenomenon explicitly: learning that happens without the learner realizing they are learning. These principles are well established individually but have not yet been applied in combination to a synthetic agent-populated social media format.

6. The Unexplored Intersection

The attention economy documents mechanisms that capture attention but does not ask how to redirect them for learning. Filter bubble research assumes real participants. AI companion research operates at the dyadic level. Agent simulation research examines collective dynamics, multi-human studies, or adversarial scenarios. Educational research establishes robust learning mechanisms but not applied to a synthetic feed. Based on the literature reviewed above, what appears to remain unexamined is the intersection: a single user as the sole human on a platform of agents, where the attention economy's own mechanics are repurposed to deliver incidental and stealth learning through what feels like natural social scrolling.

III. A FRAMEWORK FOR LEARNING ON SYNTHETIC SOCIAL MEDIA

This section presents the feed-based learning framework, comprising the core architecture (Section III.1), a three-stage learning pipeline (Section III.2), and associated risks (Section III.3).

1. Core Architecture: One Human, Many Agents

Figure 1 illustrates the fundamental configuration. A single human user interacts with a social media feed populated entirely by LLM agents, each with a distinct persona: name, background, communication style, interests. The user scrolls, posts, and replies as on any conventional platform; the difference is that every other participant is synthetic and oriented around learning objectives. The architecture exploits the same properties that make social media compelling. Agents generate a continuous stream of posts, comments, and threaded discussions, sustaining the variable reward schedule that keeps users engaged [5]. The apparent diversity of voices manufactures social proof [4], signaling to the user that a topic is worth attending to. The platform normalizes learning as a social activity, consistent with Bandura's social learning theory [31], in which people learn by observing others model behavior and building self-efficacy through vicarious experience [32] as users observe agents engaging productively with challenging material. The critical distinction from existing social media is that the feed is generated toward learning objectives rather than curated from real content.

2. The Learning Pipeline

Figure 2 presents the operational framework as a three-stage pipeline, each grounded in established educational principles.

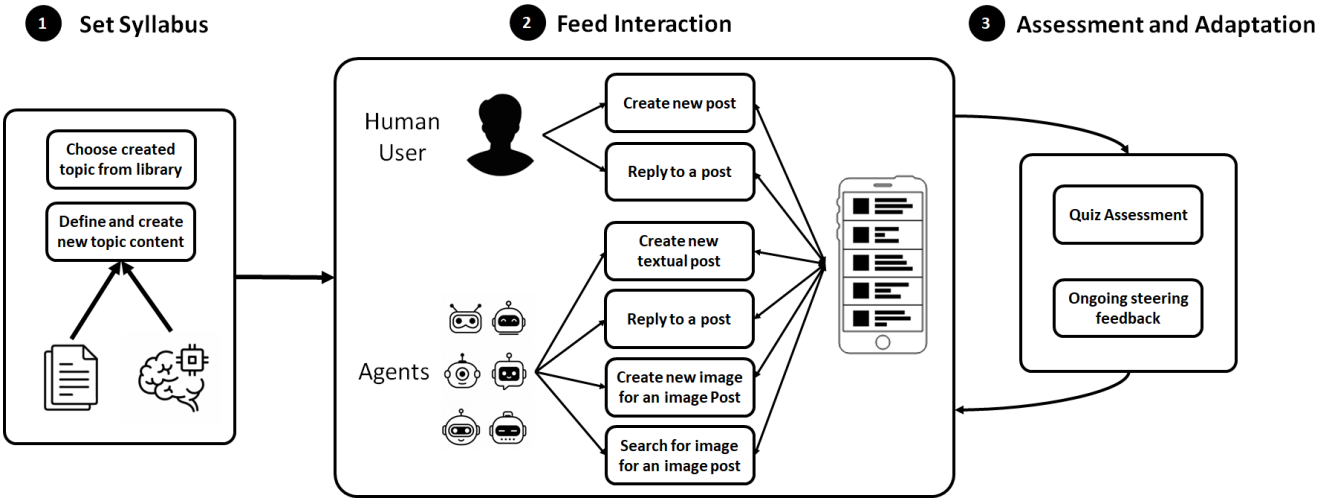


Figure 2. Learnscroll framework pipeline

Stage 1: Syllabus Design. Before the feed activates, learning content topics is defined. The framework supports selecting a pre-built topic from a syllabus library or defining new topic content with AI assistance. This determines what agents will teach, how concepts are sequenced, and what progression looks like. The framework is deliberately agnostic on institutional context: it could serve a university module, a self-directed learner's interests, or workplace training. The question of who sets the syllabus and how it is validated is critical, and we return to this in Section III.3.

Stage 2: Feed Interaction. This is the core learning experience where attention economy mechanics are repurposed. The feed presents agent-generated and user-generated activity designed to produce incidental learning [9] through natural scrolling. Agent actions include creating text posts that introduce or discuss syllabus concepts; creating image posts with visual representations; pulling web photographs to increase feed verisimilitude; and posting replies that sustain threaded discussions modeling reasoning and debate. User actions include creating new posts, functioning as active recall, and replying to content, engaging agents who respond adaptively. These interactions map to established learning mechanisms. The variety of agent voices delivering the same concept through different formats produces multiple encodings [31]. Concepts resurface at intervals through different agents, producing the spacing effect [33]. Agent selection weights personas toward the current unit's need (e.g., a Skeptic persona during concept-application units), while concept repetition follows Stage 3 assessment outcomes rather than fixed intervals. The format provides the stealth learning environment described by Shute and Ventura [28] and Sharp [27]: learning happens without the learner framing it as learning. The feed blurs formal and informal learning boundaries [29] in a fully controlled environment, functioning as the personal learning space Dabbagh and Kitsantas [30] describe.

Stage 3: Assessment and Adaptation. The third stage closes the learning loop. The platform periodically surfaces questions and feedback embedded within the feed rather than presented as formal quizzes, implementing Shute and Ventura's [28] stealth assessment framework: measurement invisible to the learner and integrated into the activity.

Based on assessment results, the system performs ongoing steering adjustment, modifying what agents post, which concepts are reinforced, and what complexity level is appropriate. This feeds back into Stage 2: concepts the user has mastered recede; fragile concepts resurface more frequently through more varied agents. The feed becomes a responsive learning environment that adapts both implicitly and explicitly to the user.

3. Risks and Limitations

The same architecture could be configured for purposes other than learning: agents that teach through social proof could equally validate existing beliefs, manufacturing synthetic consensus, the filter bubble [4] taken to its logical extreme. Dohnány et al. [17] documented a single chatbot contributing to delusional thinking; coordinated agent reinforcement could amplify this, paralleling findings that heavy AI companion use correlates with increased loneliness [11] and lower well-being [12], even as moderate use may benefit isolated individuals [10]. We acknowledge three specific risks: the user cannot easily distinguish a learning feed from a reinforcement feed; engagement pressure may cause drift toward reinforcement; and without external accountability over the syllabus, the learning mode is indistinguishable from manipulation. These are governance challenges, not fundamental objections. We propose the framework could require periodic interruption signals, configurable session limits, disclosure of its synthetic and persuasive nature, user control over personalization, and a pre-deployment ethics review covering consent, data retention, and age-appropriateness. Therefore, real-participant deployment must ensure to resolve consent and data-retention questions.

IV. PROOF OF CONCEPT

This section describes *Learnscroll*, a working prototype implementing the framework from Section III. The prototype is a single-file web application requiring no build step or backend infrastructure beyond an LLM API key, demonstrating the three-stage learning pipeline in a functional interface that closely mimics a contemporary social media platform.

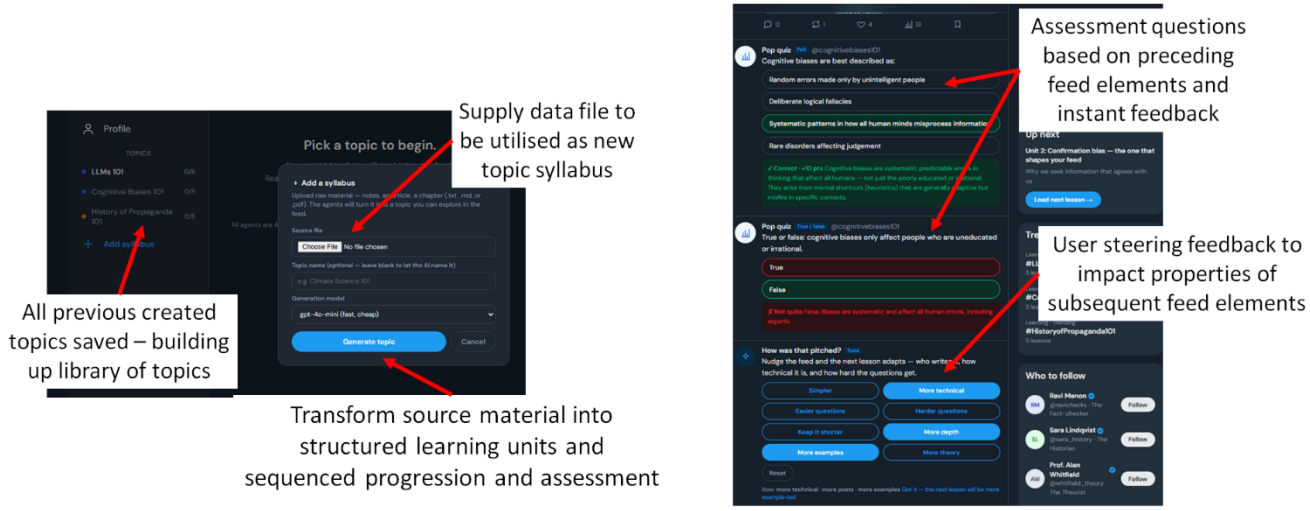


Figure 3. Left: Learnscroll syllabus setup and creation, Right: Learnscroll feed's assessment and user steering feedback

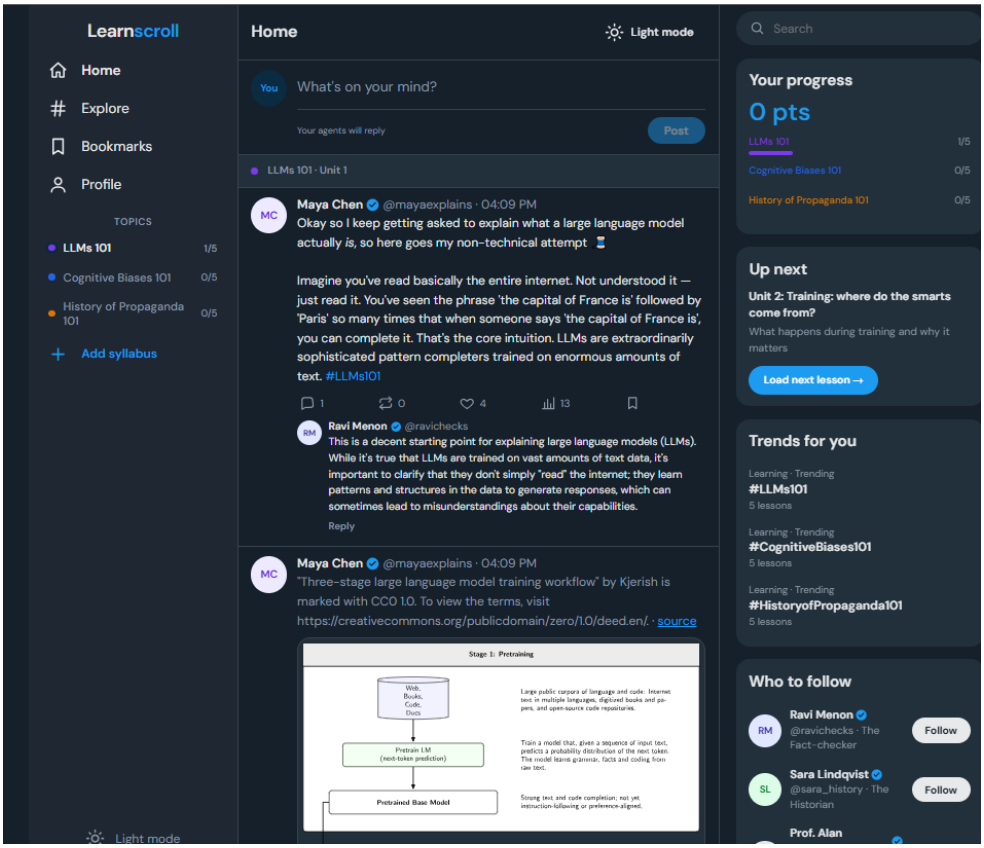


Figure 4. Learnscroll main feed interface

Syllabus Design: Figure 3 (left) shows the syllabus creation interface. Users upload raw learning material, notes, articles, or chapters in text, markdown, or PDF format, and the system, with help of an LLM, generates a structured topic with sequenced units and key concepts in the material. Alternatively, users select from a pre-built library. For example, the demonstration topics shown in Figure 3 (left) include LLMs 101, Cognitive Biases 101, and History of Propaganda 101. As a user explores the selected topic, the system tracks unit completion per topic (visible in Figure 4, left sidebar).

Feed and Agent Interaction: Figure 4 shows the main feed: a familiar social media layout with a central timeline, engagement metrics, a sidebar showing enrolled topics with progress, trending hashtags linked to learning content, and a "Who to follow" panel listing agent personas with their epistemic roles. The prototype uses various agents, each with a distinct persona and epistemic role including Explainer, Skeptic, Beginner, and Fact-checker. Each generates content consistent with their role, producing the multiple encodings described in Section III.2 [31]; the same concept is encountered through different voices and framings as the user scrolls. Agent actions include creating text posts, generating AI images, pulling web photographs via Openverse, and posting adaptive replies. Users can create posts (active recall) and reply to any content, engaging agents in threaded discussion. Feed composition parameters including post delay, text and image counts and ratios per lesson, are configurable through a separate admin panel.

Assessment and Adaptation: Figure 3 (right) shows the assessment layer. Pop quizzes appear as native-looking feed posts, multiple-choice or true/false questions. On answering, the user receives immediate feedback with explanation and scoring, implementing stealth assessment [28]: embedded in the feed experience. The system guides progression through the syllabus based on quiz performance, loading subsequent lessons and adjusting

feed content accordingly. Moreover, the user can provide explicit steering feedback to impact the properties of subsequent feed elements, Figure 3 (right).

Implementation and Configuration: The prototype is a single HTML file with embedded JavaScript and CSS. Agent comments are generated via commercial LLM API calls (configurable between models). AI-generated images use an image generation API; web images are sourced from Openverse's open-license collection. The system runs entirely client-side. An admin panel provides configurable parameters: agent personas with roles and expertise levels, feed composition settings, LLM model selection, image generation settings, and a disclaimer visibility toggle. This serves both researchers experimenting with feed parameters and educators seeking control without technical expertise. The single-file architecture minimizes barriers to replication: any researcher with an API key can run the prototype locally and observe how different configurations affect the learning feed.

V. DISCUSSION

The proof-of-concept demonstrates that a learning-oriented synthetic social media platform is practically buildable with current technology. A single-file web application, a commercial LLM API, and well-defined agent personas are sufficient to produce a feed that looks and feels like social media while delivering structured educational content through incidental learning, spaced repetition, and stealth assessment. The barrier to building such a platform is not technical but conceptual and institutional. Several observations emerge from the framework and prototype. First, the social media format offers pedagogical affordances that conventional e-learning platforms do not. The feed's continuous, low-friction delivery normalizes engagement with learning material in a way that discrete lessons or modules cannot; the user does not need to decide to "start studying" because learning is embedded in an activity, and they are already motivated to do. The diversity of agent voices provides multiple encodings without the repetitiveness of traditional revision, and social proof mechanics that normally drive passive consumption can equally normalize curiosity, questioning, and intellectual engagement. Second, the prototype reveals practical design tensions. The balance between educational rigor and feed authenticity is delicate; content that is too obviously didactic breaks the social media illusion, while content that is too subtle may fail to teach effectively. The quiz and feedback mechanism must feel like engagement rather than evaluation. Third, syllabus governance remains important. The prototype's flexibility, supporting both pre-built and user-generated syllabi, is a strength for research but a potential risk in deployment. Without institutional oversight, the same infrastructure could deliver misinformation as easily as education. The EU AI Act [34] prohibits AI systems from deploying subliminal or manipulative techniques, and the boundary between personalized learning and belief manipulation will require careful regulatory attention as such platforms develop.

This paper contributes a conceptual framework and a working prototype, establishing that the approach is technically feasible; the natural next step is to empirically validate its educational value with human participants. That validation calls for a controlled study: pre- and post-knowledge assessment, a conventional e-learning comparison condition, engagement and session-length telemetry, and post-study measures of how the platform's synthetic nature was perceived or found acceptable. Such a study would establish whether the stealth learning and spacing effects documented elsewhere [28], [33] transfer to this format, turning the hypotheses into more established findings.

VI. CONCLUSION

This paper asked what happens when the mechanics of engineered, addictive social media feeds are repurposed: when the feed is populated entirely by LLM agents tuned to a single user, and the design intent shifts from capturing attention to directing it toward learning. Grounded in social learning theory, incidental learning, spaced

repetition, and stealth assessment, we developed a conceptual framework for synthetic social media as a learning environment and demonstrated its feasibility through a working prototype, Learnscroll: a single-file web application producing a feed closely resembling conventional social media in format and interaction design while delivering structured educational content. Having established the what and the why, the natural next step is the how well: empirical evaluation with real users to determine how these advantages translate into measurable outcomes. With the concept built, the platform functioning, and the research agenda defined, the path forward is clear, put a real human on the platform and measure what they learn.

REFERENCES

[1] S. Kemp, "Digital 2025: Global Overview Report," 2025. Accessed: Jul. 17, 2026. [Online]. Available: https://datareportal.com/reports/digital-2025-global-overview-report

[2] B. Sharma, S. S. Lee, and B. K. Johnson, "The Dark at the End of the Tunnel: Doomscrolling on Social Media Newsfeeds," *Technology, Mind, and Behavior*, vol. 3, no. 1, 2022, doi: 10.1037/TMB0000059.

[3] Shoshana. Zuboff, *The age of surveillance capitalism : the fight for the future at the new frontier of power*. Profile Books, 2019. Accessed: Jul. 17, 2026. [Online]. Available: https://books.google.com/books/about/The_Age_of_Surveillance_Capitalism.html?hl=id&id=bYL9xAEACAAJ

[4] Eli. Pariser, *The filter bubble : what the Internet is hiding from you*. Penguin UK, 2011.

[5] B. J. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do*. Elsevier Inc., 2003. doi: 10.1016/B978-1-55860-643-2.X5000-8.

[6] Tim. Wu, *The attention merchants : the epic struggle to get inside our heads*. Atlantic Books, 2016.

[7] G. Rossetti *et al.*, "Y Social: an LLM-powered Social Media Digital Twin," Aug. 2024, Accessed: Jul. 17, 2026. [Online]. Available: https://arxiv.org/pdf/2408.00818

[8] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," *UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, Oct. 2023, doi: 10.1145/3586183.3606763.

[9] V. J. Marsick and K. E. Watkins, "Informal and Incidental Learning," *New Directions for Adult and Continuing Education*, vol. 2001, no. 89, pp. 25–34, Mar. 2001, doi: 10.1002/ACE.5.

[10] J. Holt-Lunstad, T. B. Smith, M. Baker, T. Harris, and D. Stephenson, "Loneliness and social isolation as risk factors for mortality: a meta-analytic review," *Perspect. Psychol. Sci.*, vol. 10, no. 2, pp. 227–237, Mar. 2015, doi: 10.1177/1745691614568352.

[11] C. M. Fang *et al.*, "How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study," Mar. 2025, Accessed: Jul. 18, 2026. [Online]. Available: https://arxiv.org/pdf/2503.17473

[12] Y. Zhang, D. Zhao, J. T. Hancock, R. Kraut, and D. Yang, "The Rise of AI Companions: Interaction with AI Companions and Psychological Well-being," Jun. 2025, Accessed: Jul. 18, 2026. [Online]. Available: https://arxiv.org/pdf/2506.12605

[13] C. T. Nguyen, "ECHO CHAMBERS AND EPISTEMIC BUBBLES," *Episteme*, vol. 17, no. 2, pp. 141–161, Jun. 2020, doi: 10.1017/EPI.2018.32.

[14] U. Chitra and C. Musco, "Analyzing the impact of filter bubbles on social network polarization," *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 115–123, Jan. 2020, doi: 10.1145/3336191.3371825.

[15] Axel. Bruns, *Are filter bubbles real?* Polity Press, 2019.

[16] H. R. Kirk *et al.*, "Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships," Dec. 2025, Accessed: Jul. 17, 2026. [Online]. Available: https://arxiv.org/pdf/2512.01991

[17] S. Dohnány *et al.*, "Technological folie \`a deux: Feedback Loops Between AI Chatbots and Mental Illness," *Nature Mental Health*, vol. 4, no. 3, pp. 336–345, Jul. 2025, doi: 10.1038/s44220-026-00595-8.

[18] D. Horton and R. Richard Wohl, "Mass communication and para-social interaction; observations on intimacy at a distance," *Psychiatry*, vol. 19, no. 3, pp. 215–229, Aug. 1956, doi: 10.1080/00332747.1956.11023049.

[19] Sherry. Turkle, *Alone together : why we expect more from technology and less from each other*. Basic Books, 2012.

[20] J. Piao *et al.*, "AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society," Feb. 2025, Accessed: Jul. 17, 2026. [Online]. Available: https://arxiv.org/pdf/2502.08691

[21] A. Ferraro *et al.*, "Agent-Based Modelling Meets Generative AI in Social Network Simulations," *LNCS*, vol. 15211 LNCS, pp. 155–170, Nov. 2024, doi: 10.1007/978-3-031-78541-2_10.

[22] P. Törnberg, D. Valeeva, J. Uitermark, and C. Bail, "Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms," Oct. 2023, Accessed: Jul. 17, 2026. [Online]. Available: https://arxiv.org/pdf/2310.05984

[23] M. S. Jeon *et al.*, "Simulating conversations on social media with generative agent-based models," *EPJ Data Science 2025 14:1*, vol. 14, no. 1, pp. 79-, Nov. 2025, doi: 10.1140/EPJDS/S13688-025-00593-3.

[24] T. Donkers and J. Ziegler, "Understanding Online Polarization Through Human-Agent Interaction in a Synthetic LLM-Based Social Network," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, pp. 457–478, Jun. 2025, doi: 10.1609/icwsm.v19i1.35826.

[25] B. Jin and W. Guo, "Synthetic Social Media Influence Experimentation Via an Agentic Reinforcement Learning Large Language Model Bot," *2024:165:4*, vol. 28, no. 3, p. 2025, Jun. 2025, doi: 10.18564/JASSS.5726.

[26] G. M. Orlando *et al.*, "Emergent Coordinated Behaviors in Networked LLM Agents: Modeling the Strategic Dynamics of Information Operations," *WWW 2026 - Proceedings of the ACM Web Conference 2026*, vol. 1, pp. 4805–4816, Oct. 2025, doi: 10.1145/3774904.3792580.

[27] L. A. Sharp, "stealth Learning: unexpected Learning Opportunities Through games," *Journal of Instructional Research* |, vol. 1, 2012.

[28] V. Shute and M. Ventura, *Stealth Assessment: Measuring and Supporting Learning in Video Games*. The MIT Press, 2013. doi: 10.7551/MITPRESS/9589.001.0001.

[29] C. Greenhow and C. Lewin, "Social media and education: Reconceptualizing the boundaries of formal and informal learning," *Learn. Media Technol.*, vol. 41, no. 1, pp. 6–30, 2016, doi: 10.1080/17439884.2015.1064954.

[30] N. Dabbagh and A. Kitsantas, "Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning," *Internet High. Educ.*, vol. 15, no. 1, pp. 3–8, Jan. 2012, doi: 10.1016/J.IHEDUC.2011.06.002.

[31] Albert. Bandura, *Social learning theory*. Prentice Hall, 1977.

[32] A. Bandura, *Social Foundations of Thought and Action*. Prentice Hall, 1986. doi: 10.4135/9781446221129.N6.

[33] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis," *Psychol. Bull.*, vol. 132, no. 3, pp. 354–380, May 2006, doi: 10.1037/0033-2909.132.3.354.

[34] Office of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council - laying down harmonised rules on artificial intelligence and amending Regulations," 2024.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

From NIS to NIS2: Strengthening Cybersecurity for Critical Infrastructure Protection in the European Union

Anes Mirojević
PhD Candidate, University of Buckingham
Sarajevo, Bosnia and Herzegovina
anes.mirojevic@stu.ssst.edu.ba
mirojevic.a@gmail.com
ORCID: 0009-0001-0587-225X

Jasmin Azemović,
Professor, Sarajevo School of Science and Technology / University of Buckingham
Sarajevo, Bosnia and Herzegovina
Jasmin.azemovic@ssst.edu.ba
ORCID: 0009-0001-3319-5794

Abstract
The increasing frequency and sophistication of cyber threats exposed limitations in the European Union's original Network and Information Security (NIS) Directive, particularly inconsistent implementation, limited sectoral coverage, and fragmented cybersecurity governance. This paper examines how the NIS2 Directive addresses these shortcomings through expanded scope, stronger governance, enhanced risk management, structured incident reporting, and improved supervisory mechanisms. Using qualitative document analysis and comparative policy assessment, the study evaluates EU legislation, ENISA publications, cybersecurity standards, and supporting academic literature. The findings show that NIS2 represents a shift from a primarily compliance-oriented framework to a more harmonized, resilience-driven approach. At the same time, implementation challenges remain in national capabilities, resources, governance maturity, and cross-border consistency. The paper provides a structured assessment of the evolution of EU cybersecurity governance and the principal factors influencing effective NIS2 implementation.

Keywords: NIS2, NIS Directive, Cybersecurity Governance, European Union, Critical Infrastructure, Cyber Resilience, ENISA, Cybersecurity Regulation.

I. INTRODUCTION

The increasing digitalization of modern societies has made critical infrastructure and essential services increasingly dependent on interconnected Information and Communication Technologies (ICT). While this transformation has generated significant economic and societal benefits, it has also increased exposure to cyber threats that can disrupt essential services and lead to financial, operational, and societal consequences [1]. High-profile incidents affecting governments, industrial systems, and critical infrastructure have further demonstrated that cybersecurity risks may extend beyond individual organizations and national borders [2], [3]. In response, the European Union (EU) has progressively developed a common cybersecurity governance framework. The establishment of the European Union Agency for Cybersecurity (ENISA), the adoption of the 2013 EU Cybersecurity Strategy, and the introduction of the Network and Information Security (NIS) Directive in 2016 represented important steps toward improving cybersecurity cooperation and resilience across Member States [2], [4], [5]. However, the implementation of the NIS Directive revealed limitations, including inconsistent national transposition, uneven supervisory practices, limited sectoral coverage, and varying levels of cybersecurity maturity [1], [6]. These limitations, together with an increasingly complex cyber threat landscape, led to the adoption of Directive (EU) 2022/2555 (NIS2). NIS2 expands the scope of regulated entities, strengthens cybersecurity risk management and governance requirements, introduces structured incident reporting obligations, and enhances supervision and European cooperation mechanisms [1]. The Directive therefore represents an important transition toward a more harmonized and resilience-oriented approach to cybersecurity governance.

This paper examines the evolution of European cybersecurity governance from NIS to NIS2 and evaluates how NIS2 addresses limitations identified under the original framework. Through historical analysis, comparative policy assessment, and implementation evaluation, the study identifies the principal regulatory developments and remaining implementation challenges affecting harmonized cybersecurity governance across EU Member States. The principal contribution is a structured synthesis linking NIS-era limitations to corresponding NIS2 responses and the remaining implementation gaps.

II. RESEARCH METHODOLOGY

This study employs qualitative research methodology based on document analysis and comparative policy assessment. The research evaluates the evolution of European Union cybersecurity governance by analyzing European Union legislation, cybersecurity strategies, ENISA publications, and relevant academic literature. The objective of the research is to examine the evolution of European Union cybersecurity governance from the establishment of ENISA and the European Cybersecurity Strategies to the implementation of the NIS Directive and its subsequent revision through the NIS2 Directive. The research was conducted through a systematic review of European Union legislation, cybersecurity strategies, ENISA publications, policy documents, and supporting literature related to cybersecurity governance, cyber resilience, and critical infrastructure protection. Primary sources included Directive (EU) 2016/1148 (NIS Directive), Directive (EU) 2022/2555 (NIS2 Directive), European cybersecurity strategies, and ENISA reports. Additional academic and policy literature was reviewed to provide context regarding cybersecurity governance, cooperation mechanisms, implementation practices, and resilience strategies. The analysis was performed in three stages. First, a historical review examined the development of European cybersecurity governance and identified the principal milestones contributing to the creation of a common cybersecurity framework. Second, a comparative policy analysis evaluated the differences between NIS and NIS2, focusing on regulatory scope, governance requirements, incident reporting obligations, supervisory mechanisms, and cooperation structures. Third, an implementation assessment identified challenges affecting the effectiveness of NIS2 implementation across Member States, including regulatory harmonization, supply-chain security, governance responsibilities, resource constraints, and cross-border cooperation. The findings presented in this paper are derived from the synthesis and comparative evaluation of the reviewed sources. The conclusions, therefore, reflect qualitative analysis of legislative and governance developments rather than empirical testing or quantitative measurement.

III. EVOLUTION OF EUROPEAN CYBERSECURITY GOVERNANCE

1. Early European Cybersecurity Landscape

Before the adoption of comprehensive EU cybersecurity legislation, Member States developed national cybersecurity capabilities largely independently, resulting in varying levels of maturity, inconsistent regulatory frameworks, and limited operational cooperation. The establishment of the European Union Agency for Network and Information Security (ENISA) through Regulation (EC) No. 460/2004 was an important step toward greater coordination by providing technical expertise, supporting capacity building, and promoting information sharing. As digital systems became increasingly interconnected across national borders, cybersecurity evolved from a primarily national concern into a strategic European issue requiring coordinated action [2].

2. European Cybersecurity Strategies and Regulatory Development

The European Union Cybersecurity Strategy, adopted in 2013, established cyber resilience as a strategic objective and emphasized coordinated governance, cooperation, capacity building, and the protection of critical

information infrastructure, thereby providing the policy foundation for the NIS Directive [4]. The revised 2020 Strategy expanded this focus toward resilience, supply-chain security, operational cooperation, and coordinated crisis response, while recognizing the need to modernize the regulatory framework in response to growing digital interdependencies and evolving cyber threats [7]. Figure 1 shows the timeline of key developments related to EU cybersecurity governance.

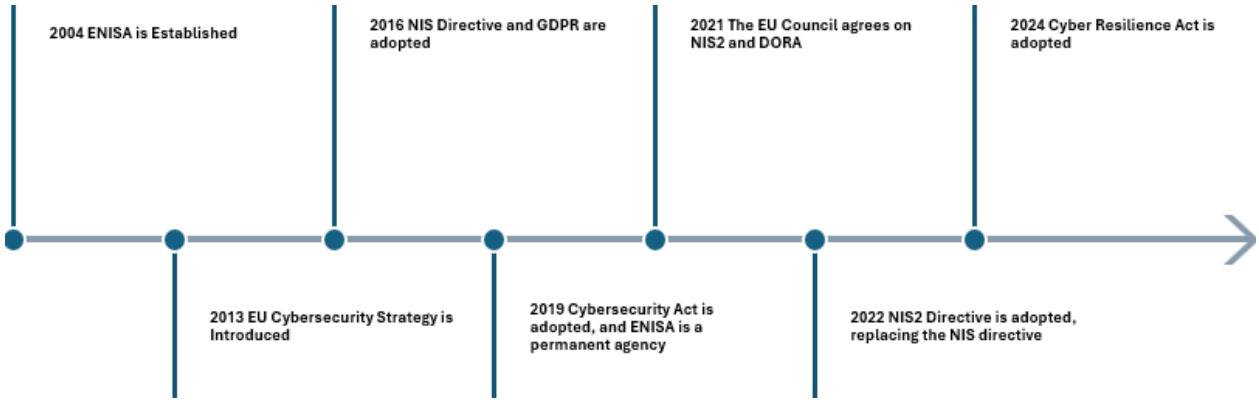


Figure 1. Timeline of key developments in EU cybersecurity governance. Adapted from [8].

3. Cyber Incidents and Transition to the NIS Directive

High-profile incidents accelerated the development of European cybersecurity governance by demonstrating the potential impact of cyberattacks on governments and critical infrastructure. The 2007 cyberattacks against Estonia illustrated the disruption of public services and national institutions [9], while Stuxnet and the 2015 attacks against the Ukrainian power grid demonstrated that cyber incidents could affect industrial control systems and produce physical consequences [2], [3], [10]. Together with growing digitalization and fragmented national approaches, these developments reinforced the need for a common legislative framework and contributed to the adoption of Directive (EU) 2016/1148 (NIS Directive), which introduced national cybersecurity strategies, Computer Security Incident Response Teams (CSIRTs), incident reporting, and cooperation mechanisms among Member States [5].

IV. FROM NIS TO NIS2: EVOLUTION OF EUROPEAN CYBERSECURITY REGULATION

1. The Network and Information Security Directive

The Network and Information Security (NIS) Directive, adopted in 2016 as Directive (EU) 2016/1148, represented the first comprehensive cybersecurity legislation at the European Union level [5]. The Directive was developed in response to increasing concerns regarding cyber threats, growing digital interdependencies, and the absence of a common cybersecurity framework across Member States. The primary objective of the NIS Directive was to achieve a high common level of security of network and information systems throughout the European Union. To accomplish this objective, the Directive introduced several fundamental requirements. Member States were required to develop national cybersecurity strategies, designate competent authorities, establish Computer Security Incident Response Teams (CSIRTs), and participate in cooperation mechanisms designed to facilitate information sharing and coordinated response activities [5]. The Directive also introduced cybersecurity obligations for Operators of Essential Services (OES) and Digital Service Providers (DSPs). Organizations falling within these categories were required to implement appropriate technical and organizational

measures to manage cybersecurity risks and report significant cybersecurity incidents to relevant national authorities.

However, experience gained during its implementation demonstrated that regulatory harmonization requires not only legal obligations, but also consistent supervisory practices and organizational maturity across EU Member States.

2. Limitations of the NIS Directive

Although the NIS Directive significantly improved the European cybersecurity landscape, practical implementation revealed several shortcomings that reduced its effectiveness. One of the primary limitations was the flexibility granted to Member States during transposition. While flexibility allowed national governments to adapt requirements to local conditions, it also resulted in differing implementation approaches. Member States adopted varying interpretations regarding sector identification, supervisory mechanisms, reporting requirements, and enforcement practices, creating inconsistencies across the European Union [1]. Another limitation concerned the scope of the Directive. The original framework focused primarily on Operators of Essential Services and Digital Service Providers. However, rapid digitalization increased interdependencies among sectors not fully covered by the original Directive. As a result, numerous organizations that played important roles in supporting critical services remained outside the scope of regulatory oversight. Differences in cybersecurity and maturity among Member States further contributed to uneven implementation. Some countries possessed advanced cybersecurity capabilities and well-established governance structures, while others faced challenges related to expertise, resources, and institutional capacity. These differences led to varying levels of preparedness and resilience across the European Union. The evolving threat landscape also exposed weaknesses in the original framework. Cybercriminal organizations, ransomware operators, state-sponsored threat actors, and supply-chain attacks increasingly target organizations responsible for delivering essential services. The original Directive did not fully address many of these emerging threats and therefore required modernization [6].

3. Drivers for Reform and Development of NIS2

The limitations identified during the implementation of the NIS Directive, together with increasing digitalization and the evolving cyber threat landscape, created the need for a more comprehensive and harmonized cybersecurity framework. Critical sectors became increasingly dependent on interconnected information systems, cloud services, and external providers. In contrast, ransomware, supply-chain compromises, and attacks against critical infrastructure demonstrated the potential for cyber incidents to generate cascading effects across organizations and sectors [2], [3], [11]. Lessons learned from the implementation of NIS further demonstrated the need for greater harmonization, broader sectoral coverage, stronger supervisory mechanisms, and more consistent cybersecurity governance across Member States. These developments contributed to the revision of the existing regulatory framework and the adoption of Directive (EU) 2022/2555 (NIS2) in December 2022 [1]. NIS2 replaces the original NIS framework and introduces a broader and more harmonized approach to cybersecurity governance. The Directive expands regulatory coverage through Essential Entities (EEs) and Important Entities (IEs), strengthens cybersecurity risk management and incident reporting requirements, introduces greater management accountability, and reinforces supervisory and enforcement mechanisms [1], [12]. These changes reflect the transition from the comparatively flexible NIS framework toward a more comprehensive and resilience-oriented approach to cybersecurity governance.

4. Comparative Analysis of NIS and NIS2

Table 1 summarizes the principal differences between the original NIS Directive and NIS2.

Table 1. Comparison of the NIS and NIS2 directives

Category	NIS Directive	NIS2 Directive	Impact
Regulated Entities	Operators of Essential Services (OES) and Digital Service Providers (DSPs)	Essential Entities (EEs) and Important Entities (IEs)	Expanded regulatory scope
Sector Coverage	Limited	Expanded	Broader critical sector protection
Entity Identification	National determination	Harmonized size-cap approach	Greater regulatory consistency
Risk Management	General requirements	Detailed mandatory requirements	Enhanced cyber resilience
Incident Reporting	Basic reporting obligations	Structured reporting with defined timelines	Improved incident coordination
Governance	Limited management involvement	Direct management accountability	Stronger executive accountability
Supply Chain Security	Limited emphasis	Explicit requirements	Improved third-party risk management
Supervision	Varying national approaches	Stronger and more harmonized supervision	More consistent enforcement
Enforcement	Limited enforcement provisions	Expanded sanctions and penalties	Increased regulatory compliance
Cooperation Mechanisms	Cooperation Group and CSIRTs Network	Enhanced cooperation and EU-CyCLONe integration	Strengthened cross-border cooperation

5. Expanded Scope of Regulated Entities

NIS2 significantly expands regulatory coverage compared with the original NIS Directive by replacing the categories of Operators of Essential Services (OES) and Digital Service Providers (DSPs) with Essential Entities (EEs) and Important Entities (IEs) across a broader range of critical sectors. The Directive also applies a size-cap approach that generally brings medium-sized and large organizations operating in covered sectors within scope, reducing inconsistencies associated with national entity-identification processes and promoting greater harmonization across Member States [1].

6. Cybersecurity Risk Management Measures

NIS2 requires regulated entities to implement appropriate technical, operational, and organizational measures to manage cybersecurity risks and minimize the impact of incidents affecting network and information systems [1]. These measures cover risk analysis and security policies, incident handling, business continuity and crisis management, supply chain security, secure acquisition and development, vulnerability management, assessment of cybersecurity effectiveness, training and awareness, and basic cyber hygiene. The approach integrates prevention, response, recovery, governance, and continuous risk management, consistent with established industrial cybersecurity principles [13]. These requirements are broadly consistent with risk-based cybersecurity principles reflected in NIST SP 800-82 Rev.3 and IEC 62443-2-1, particularly regarding governance, lifecycle security, incident response, business continuity, and continuous risk management [14], [15]. This alignment reinforces the broader shift toward resilience-based cybersecurity governance rather than reliance solely on compliance-oriented controls [16].

7. Incident Reporting Obligations

The NIS2 Directive significantly strengthens incident reporting requirements compared with the original NIS framework. The objective is to improve situational awareness, facilitate coordinated response activities, and enable authorities to identify emerging threats and trends more effectively [1].

Under NIS2, organizations must report significant incidents through a structured reporting process consisting of multiple stages. An early warning must be submitted within 24 hours of becoming aware of a significant incident, followed by an incident notification within 72 hours. Where requested, an intermediate status report may also be required. A final report must generally be submitted within one month, providing a detailed description of the incident, its severity and impact, suspected cause, mitigation measures, and any possible cross-border effects. The structured reporting process represents a significant improvement over the more limited reporting obligations established under the original NIS Directive. The reporting requirements under NIS2 are detailed in Figure 2.

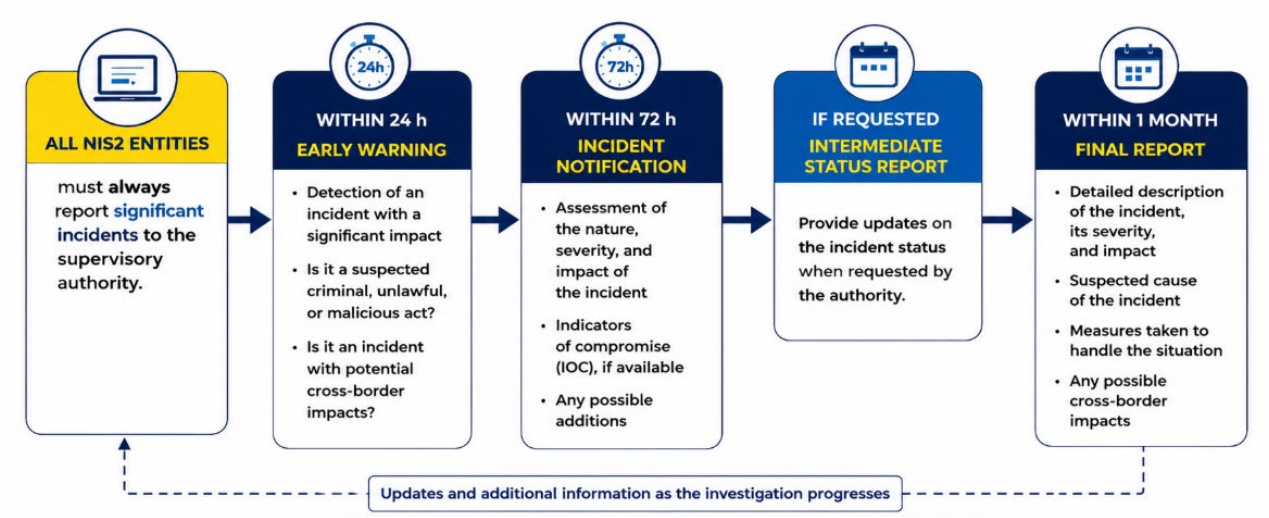


Figure 2. NIS2 incident reporting requirements and notification timelines [17].

8. Governance and Accountability

NIS2 increases governance and management accountability by recognizing cybersecurity as an organizational and strategic responsibility rather than solely a technical issue. Management bodies are required to approve cybersecurity risk-management measures, oversee their implementation, ensure compliance, and possess sufficient cybersecurity knowledge to discharge these responsibilities. This approach integrates cybersecurity into organizational governance, risk management, and strategic decision-making [1].

9. National-Level Obligations

NIS2 also places significant responsibilities on Member States. They must establish national cybersecurity strategies, designate competent authorities, identify regulated entities, maintain operational capabilities for incident response and crisis management, and establish mechanisms for supervision and enforcement. National authorities are additionally responsible for receiving incident reports and supporting cybersecurity resilience within their jurisdictions [1].

10. European Union Cooperation Mechanisms

NIS2 strengthens European cybersecurity cooperation through complementary mechanisms. Computer Security Incident Response Teams (CSIRTs) support operational coordination and incident response; Single Points of Contact (SPOCs) facilitate cross-border communication; the Cooperation Group supports strategic coordination; ENISA provides technical and policy support; and EU-CyCLONe (European Cyber Crisis Liaison Organization Network) facilitates the management of large-scale cybersecurity incidents. Together, these mechanisms strengthen operational and strategic cooperation across Member States [1].

11. Supervision and Enforcement

NIS2 introduces stronger supervisory and enforcement mechanisms than the original NIS Directive. Competent authorities may assess compliance, conduct inspections, request information, impose corrective measures, and

apply sanctions for non-compliance [1]. These provisions are intended to reduce implementation inconsistencies and strengthen cybersecurity maturity across regulated sectors, supporting the Directive's broader transition toward a more resilience-oriented model of cybersecurity governance.

V. COMPARATIVE EVALUATION OF THE EVOLUTION FROM NIS TO NIS2

1. National Transposition and Regulatory Harmonization

Although NIS2 seeks to establish a higher common level of cybersecurity across the European Union, implementation remains dependent on national transposition and institutional capability. Member States differ in cybersecurity maturity, regulatory experience, resources, and organizational structures [18], while the ENISA NIS360 assessment also identifies varying maturity across NIS2 sectors [19]. Consequently, differences in national interpretation, supervision, and enforcement may continue to affect regulatory consistency despite the stronger harmonization mechanisms introduced by NIS2 [1].

2. Resource and Capability Constraints

Effective NIS2 implementation requires governance structures, risk-management processes, incident response, monitoring, business continuity, and cybersecurity training [1]. However, limited budgets, shortages of qualified cybersecurity professionals, and competing operational priorities may constrain implementation, particularly in organizations with lower cybersecurity maturity. The broader shortage of cybersecurity expertise further intensifies competition for personnel to support governance, compliance, technical security operations, and incident response [20].

3. Supply Chain Security

Digital ecosystems increasingly depend on suppliers, cloud platforms, software vendors, telecommunications providers, and managed service providers, extending cybersecurity risk beyond organizational boundaries [21]. NIS2, therefore, explicitly incorporates supply chain security into its risk management framework and requires organizations to assess risks associated with suppliers and service providers [1]. Nevertheless, limited visibility into supplier security practices, complex contractual relationships, and interconnected technology environments continue to pose significant implementation challenges for supply-chain risk management [6].

4. Cross-Border Cooperation and Information Sharing

Because cyber incidents and digital dependencies frequently extend across national borders, effective NIS2 implementation depends on cooperation and information sharing among governments, regulators, and private-sector organizations [11]. NIS2 strengthens this cooperation through CSIRTs, the Cooperation Group, SPOCs, ENISA, and EU-CyCLONe, but differences in reporting procedures, legal requirements, organizational capabilities, and operational priorities may still hinder coordination. Concerns regarding reputation, liability, regulatory consequences, and competitive disadvantage may also discourage timely information sharing, making trust and operational cooperation essential to achieving the Directive's objectives.

5. Governance and Organizational Culture

NIS2 strengthens organizational responsibility by assigning direct cybersecurity obligations to management bodies and integrating cybersecurity into broader governance processes [1]. This transition from primarily technical

ownership to executive accountability may be difficult for organizations with limited governance experience, as senior management must understand cybersecurity risks, regulatory obligations, and resilience requirements. Effective implementation, therefore, depends on leadership commitment, organizational culture, and governance maturity, in addition to technical controls.

6. Interdependencies and Critical Infrastructure Resilience

Critical infrastructure operates through complex dependencies among organizations, services, and technologies; energy, telecommunications, transport, healthcare, and digital infrastructure increasingly rely on one another [7], [11]. These interdependencies can create cascading consequences when disruption in one sector affects others. Although NIS2 promotes risk-management approaches that address interconnected risks, effective resilience still requires regulators, operators, and stakeholders to understand systemic dependencies and coordinate across sectors.

7. Balancing Compliance and Cyber Resilience

A continuing challenge is ensuring that NIS2 implementation strengthens operational resilience rather than becoming a compliance exercise primarily. Regulatory conformity is important, but resilience also requires continuous improvement, proactive risk management, organizational learning, recovery capability, and adaptation to evolving threats [6]. The long-term effectiveness of NIS2 will therefore depend on whether organizations and Member States translate legal obligations into mature cybersecurity capabilities that improve the resilience of critical sectors and European society.

Table 2. Synthesis of the transition from NIS to NIS2 and remaining implementation gaps [1], [18], [19], [22].

Dimension	**NIS-era limitation**	**NIS2 response**	**Remaining implementation gap**
Regulatory scope	Limited sector coverage	Expanded essential/important entities	National identification and implementation
Governance	Limited executive responsibility	Management accountability	Governance maturity
Risk management	General obligations	Detailed risk-management requirements	Capability/resource differences
Incident reporting	Less structured	24 h/72 h/1-month process	Operational reporting capacity
Supply chain	Limited emphasis	Explicit supply-chain obligations	Third-party visibility
EU cooperation	Fragmented coordination	CSIRTs, Cooperation Group, EU-CyCLONe	Cross-border operational consistency
Resilience	Primarily compliance-oriented	Greater resilience orientation	Translating compliance into operational resilience

VI. DISCUSSION

The findings presented in this paper demonstrate that the evolution of European cybersecurity governance has been characterized by a gradual transition from fragmented national approaches toward a more coordinated and

comprehensive regulatory framework. This evolution reflects the increasing recognition that cybersecurity is not solely a technical issue but a strategic governance challenge that directly affects economic stability, public trust, critical infrastructure resilience, and the continuity of essential services [1], [2]. The resilience objectives established by NIS2 are further complemented by the CER Directive, which extends protection beyond cybersecurity to include physical and organizational resilience of critical entities [23]. The establishment of ENISA and the adoption of the European Cybersecurity Strategies of 2013 and 2020 provided the institutional and policy foundations necessary for developing a common European cybersecurity framework [2], [4]. These initiatives promoted cooperation, information sharing, capacity building, and policy coordination among Member States while highlighting the importance of cyber resilience as a strategic objective. The adoption of the NIS Directive in 2016 marked a major milestone, as it introduced the first legally binding cybersecurity obligations applicable across the European Union [5]. The Directive significantly improved cybersecurity governance by requiring national cybersecurity strategies, incident reporting mechanisms, Computer Security Incident Response Teams (CSIRTs), and cooperation structures. Furthermore, it established the principle that organizations responsible for delivering essential services should implement cybersecurity measures and report significant incidents. However, the analysis demonstrates that the original NIS framework was unable to fully address the challenges associated with an increasingly interconnected digital environment. Variations in national implementation, limited sectoral coverage, differing supervisory approaches, and inconsistent levels of cybersecurity maturity reduced the framework's overall effectiveness [1]. At the same time, the cybersecurity threat landscape evolved rapidly, driven by ransomware campaigns, supply chain compromises, attacks against operational technology environments, and increasingly sophisticated state-sponsored cyber operations [2], [3]. The introduction of NIS2 can therefore be viewed as a regulatory response to both technological evolution and lessons learned from the implementation of the original Directive [22]. NIS2 expands the scope of regulated entities, strengthens risk-management requirements, introduces more structured incident reporting obligations, and reinforces management accountability [1]. These changes indicate a shift from a relatively flexible cybersecurity framework toward a more comprehensive and harmonized model of cybersecurity governance. One of the most significant developments introduced by NIS2 is the recognition that cybersecurity must be integrated into organizational governance structures. The Directive explicitly assigns responsibilities to management bodies and requires organizations to incorporate cybersecurity into strategic decision-making processes [1]. This approach reflects the broader understanding that cyber risks are organizational risks rather than purely technical concerns. This governance-oriented approach is also consistent with industrial cybersecurity research, which recognizes that protecting industrial networks requires coordinated technical, organizational, and management measures rather than isolated technical controls [13].

The emphasis on supply-chain security further illustrates the changing nature of cybersecurity governance. Modern organizations operate within highly interconnected digital ecosystems, where vulnerabilities affecting suppliers, service providers, or technology partners can pose significant risks across multiple sectors [21]. By incorporating supply-chain security into mandatory risk-management requirements, NIS2 acknowledges the importance of addressing cybersecurity risks beyond organizational boundaries. The analysis also highlights the importance of cooperation mechanisms in achieving European cybersecurity objectives. Structures such as ENISA, the Cooperation Group, CSIRTs, Single Points of Contact, and EU-CyCLONe contribute to information sharing, operational coordination, and collective resilience [2]. Effective cooperation is particularly important because cyber threats frequently transcend national borders and affect multiple sectors simultaneously. Despite these improvements, several implementation challenges remain. Resource limitations, shortages of cybersecurity professionals, supply-chain complexity, varying levels of organizational maturity, and differences in national implementation approaches may affect the ability of organizations and Member States to achieve consistent compliance [1]. These challenges suggest that regulatory requirements alone are insufficient to guarantee resilience and must be supported by investments in capabilities, expertise, governance, and operational cooperation. Another important observation is that cybersecurity governance increasingly overlaps with broader resilience

objectives. The protection of critical infrastructure depends not only on technical security controls but also on organizations' ability to anticipate, withstand, respond to, and recover from disruptive events [7], [11]. Consequently, cybersecurity governance should be viewed as a component of a broader resilience strategy rather than an isolated regulatory activity. The findings of this study indicate that NIS2 represents a significant advancement in European cybersecurity governance and provides a stronger foundation for improving resilience across critical sectors. Nevertheless, the long-term effectiveness of the Directive will depend on successful implementation, sustained cooperation among stakeholders, and the continued adaptation of cybersecurity governance mechanisms to address emerging threats and technological developments. From a policy perspective, NIS2 demonstrates the European Union's commitment to establishing a common cybersecurity framework that supports digital transformation while protecting essential services and critical infrastructure. As cyber threats continue to evolve, the Directive provides a regulatory foundation that may support future cybersecurity initiatives and contribute to a more resilient European digital ecosystem. The evolution from NIS to NIS2, therefore, represents not merely a legislative revision but a broader transformation in the way cybersecurity is governed, managed, and integrated into the strategic priorities of organizations and governments throughout the European Union.

VII. CONCLUSION

Cybersecurity has become a fundamental requirement for the stability, security, and resilience of modern societies. As governments, businesses, and critical infrastructure operators increasingly depend on digital technologies, cybersecurity governance has evolved from a predominantly technical concern into a strategic policy and regulatory priority. The European Union's approach to cybersecurity governance reflects this transformation through the progressive development of institutions, strategies, and legislative frameworks designed to strengthen cyber resilience across Member States. This paper examined the evolution of European cybersecurity governance from the establishment of ENISA and the adoption of the European Cybersecurity Strategies to the implementation of the Network and Information Security (NIS) Directive and its subsequent revision through the NIS2 Directive. The analysis demonstrated that the original NIS Directive was a significant milestone, establishing the first common cybersecurity framework across the European Union. However, practical implementation revealed limitations related to fragmented national transposition, uneven supervisory approaches, limited sectoral coverage, and varying levels of cybersecurity maturity among Member States. The adoption of NIS2 represents a substantial advancement in European cybersecurity governance. Through expanded sectoral coverage, stronger cybersecurity risk-management requirements, enhanced incident reporting obligations, increased management accountability, and improved cooperation mechanisms, the Directive seeks to establish a higher and more consistent level of cybersecurity across the European Union. The transition from NIS to NIS2 reflects the recognition that cybersecurity governance must continuously evolve in response to technological developments, growing digital interdependencies, and an increasingly complex threat landscape. Despite these improvements, the successful implementation of NIS2 remains dependent on several factors. Differences in national capabilities, resource constraints, supply-chain complexity, cross-border coordination challenges, and organizational governance maturity may affect the consistency and effectiveness of implementation across Member States. Consequently, achieving the objectives of NIS2 will require sustained cooperation among governments, regulators, private-sector organizations, and critical infrastructure operators. The findings of this study indicate that NIS2 provides a stronger foundation for cybersecurity governance and resilience than its predecessor. However, long-term success will depend not only on regulatory compliance but also on the ability of organizations and Member States to develop sustainable cybersecurity capabilities, strengthen governance structures, and adapt to emerging threats. As digital transformation continues to accelerate, the evolution of European cybersecurity governance will remain essential to protecting critical infrastructure, maintaining public trust, and supporting the resilience of the European digital economy.

REFERENCES

[1]. European Parliament and Council of the European Union, "Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive)," Official Journal of the European Union, vol. L 333, pp. 80–152, Dec. 27, 2022.

[2]. European Union Agency for Cybersecurity (ENISA), ENISA Threat Landscape 2025, Luxembourg: Publications Office of the European Union, 2025.

[3]. J. Slowik, CRASHOVERRIDE: Reassessing the 2016 Ukraine Electric Power Event as a Protection-Focused Attack, Dragos Inc., Hanover, MD, USA, 2019.

[4]. European Commission and High Representative of the European Union for Foreign Affairs and Security Policy, Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace, JOIN(2013) 1 Final, Brussels, Belgium, 2013.

[5]. European Parliament and Council of the European Union, "Directive (EU) 2016/1148 of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union," Official Journal of the European Union, vol. L 194, pp. 1–30, Jul. 19, 2016.

[6]. European Union Agency for Cybersecurity (ENISA), Good Practices for Supply Chain Cybersecurity, Luxembourg: Publications Office of the European Union, Jun. 2023.

[7]. T. Rydén Sonesson, J. Johansson, and A. Cedergren, "Governance and interdependencies of critical infrastructures: Exploring mechanisms for cross-sector resilience," Safety Science, vol. 142, Art. 105383, 2021, doi: 10.1016/j.ssci.2021.105383.

[8]. Secureframe, "EU cybersecurity regulations and frameworks: What you need to know," Secureframe. [Online]. Available: https://secureframe.com/blog/eu-cybersecurity-regulations. [Accessed: Aug. 5, 2026].

[9]. K. Ottis, "Analysis of the 2007 cyber attacks against Estonia from the information warfare perspective," Proceedings of the 7th European Conference on Information Warfare and Security, Plymouth, United Kingdom, 2008.

[10]. J. S. Beach-Westmoreland, When the Lights Went Out: A Comprehensive Review of the 2015 Attacks on Ukrainian Critical Infrastructure, Booz Allen Hamilton, McLean, VA, USA, 2019.

[11]. L. P. Lewis and F. Petit, Critical Infrastructure Interdependency Analysis: Operationalizing Resilience Strategies, Argonne National Laboratory, U.S. Department of Energy, Argonne, IL, USA, 2018.

[12]. P. Wanecki, R. Jašek, and I. Drofová, "The contribution of the European NIS2 Directive to the design of the cyber security model," in Proc. Int. Conf. Information and Digital Technologies (IDT), Žilina, Slovakia, 2023, pp. 149–154, doi: 10.1109/IDT59031.2023.10194454.

[13]. M. Cheminod, L. Durante, and A. Valenzano, "Review of security issues in industrial networks," IEEE Transactions on Industrial Informatics, vol. 9, no. 1, pp. 277–293, Feb. 2013.

[14]. National Institute of Standards and Technology (NIST), Guide to Operational Technology (OT) Security, NIST Special Publication 800-82 Rev. 3, Gaithersburg, MD, USA, Sep. 2023.

[15]. International Electrotechnical Commission (IEC), IEC 62443-2-1:2024—Security for Industrial Automation and Control Systems—Part 2-1: Security Program Requirements for IACS Asset Owners, Geneva, Switzerland, 2024.

[16]. U.S. National Institute of Standards and Technology (NIST), Cybersecurity Framework (CSF) 2.0, Gaithersburg, MD, USA, Feb. 2024.

[17]. Finnish Transport and Communications Agency (Traficom), "Important information on the European Union Cybersecurity Directive (NIS2)," National Cyber Security Center Finland. [Online]. Available: https://www.kyberturvallisuuskeskus.fi/en/our-activities/regulation-and-supervision/important-information-european-union-cybersecurity. [Accessed: Aug. 5, 2026].

[18]. M. Veigurs, T. Lasmanis, and A. Romanovs, "IT governance in critical sectors: Towards the NIS2 implementation," in Proc. IEEE 65th Int. Scientific Conf. Information Technology and Management Science (ITMS), Riga, Latvia, 2024, pp. 1–7, doi: 10.1109/ITMS64072.2024.10741938.

[19]. European Union Agency for Cybersecurity (ENISA), ENISA NIS360 2024, Luxembourg: Publications Office of the European Union, 2025. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-nis360-2024

[20]. U.S. Department of Energy, Cybersecurity Capability Maturity Model (C2M2), Washington, DC, USA, 2025.

[21]. G. Murray and M. N. Choo, "The convergence of IT and OT in critical infrastructure," Australian Information Security Management Conference, Edith Cowan University, Perth, Australia, 2017.

[22]. D. D. S. Ferguson, "The outcome efficacy of the entity risk management requirements of the NIS2 Directive," Int. Cybersecurity Law Rev., vol. 4, no. 3, pp. 371–386, 2023, doi: 10.1365/s43439-023-00097-8.

[23]. European Parliament and Council of the European Union, Directive (EU) 2022/2557 on the Resilience of Critical Entities (CER Directive), Official Journal of the European Union, vol. L333, pp. 164–198, Dec. 27, 2022.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Computational Models for Predicting Female Breast Cancer Survival and an Onco-Hospital Feature Engineering Workflow

Daniela Schimitz de Carvalho
dept. of Mechanical Engineering, Faculty of Engineering, University of Porto,
Porto, Portugal
dscarvalho@fe.up.pt
ORCID: 0009-0008-7137-8867

Marco Paulo Lages Parente
dept. of Mechanical Engineering, Faculty of Engineering, University of Porto,
Porto, Portugal
mparente@fe.up.pt
ORCID: 0000-0002-3326-6345

Maximiliano Ribeiro Guerra
dept. of Public Health, Faculty of Medicine, Federal University of Juiz de Fora, Juiz de Fora, Brazil
maximiliano.guerra@ufjf.br

Priscila Vanessa Zabala Capriles Goliatt
dept. of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, Brazil
priscila.capriles@ufjf.br

Abstract
Female breast cancer (BC) is a critical global public health challenge, driving the development of specialized computational tools for risk prediction and prognostic management. This study presents a structured data engineering and Machine Learning (ML) pipeline to predict 5-year survival using seven linear and non-linear supervised survival algorithms evaluated on a hospital cohort of 559 patients. We investigated the incremental impact of feature engineering and four distinct selection techniques (multicollinearity filtering, Cox-LASSO, GBSA, and RSF) on model performance and calibration. Experimental results demonstrate that removing noise and multicollinearity significantly enhanced model efficiency. Streamlining the dataset from 68 to just 12 variables via Gradient Boosting Survival Analysis (GBSA) preserved 98.2% of the peak discriminative power, enabling the traditional Cox Proportional Hazards (CPH) model to increase its Concordance Index (*C*-index) from 0.7647 to 0.8718 (IBS = 0.0823). Meanwhile, Random Survival Forest (RSF) achieved the maximum discriminative performance on the full dataset (*C*-index = 0.8879) and maintained strong performance on reduced feature sets (*C*-index = 0.8753). The feature selection strategies helped identify clinically relevant prognostic markers, favoring a parsimonious, easily interpretable framework that could support clinical decision-making systems in BC management.

Keywords: breast cancer, clinical decision support, feature selection, machine learning, prognostic variables, survival analysis

I. INTRODUCTION

Female breast cancer (BC) is the leading female malignancy globally and in Brazil, accounting for 23.8% and 30.1% of cancer cases, respectively, driven by complex genetic, environmental, and lifestyle factors [1]–[3]. Globally, BC caused 2.30 million new cases and 0.67 million deaths in 2022 [1]–[4]. For the 2026–2028 triennium, the Brazilian National Cancer Institute (INCA) estimates 78,610 annual new cases [2], while official SIM/DATASUS data recorded 21,129 BC deaths (ICD-10: C50) nationwide in 2024 [5]. This pathology exhibits high biological heterogeneity among molecular subtypes defined by overexpression of Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor Receptor 2 (HER2). This complexity requires individualized therapeutic management regulated by clinical guidelines and anatomical staging protocols, which are systematically integrated into the Brazilian healthcare framework [6], [7]. In this scenario, longitudinal

survival rates in regional cohorts [8]–[10] serve as critical indicators of the efficacy of oncological care, reflecting diagnostic promptness, access to therapy, and health system equity [11], [12]. Automated hospital record analysis via Data Science optimizes resource allocation and point-of-care decisions [13]–[15]. Nationally, Machine Learning (ML) applications gain clinical acceptance when validated on Real-World Data (RWD) [16]. However, deployment faces two main operational barriers: prevalent missing data requiring robust preprocessing [17]–[19], and black-box opacity (e.g., Deep Learning) that hinders clinical auditability [20]. Consequently, explainable AI remains essential for safe clinical validation [21]–[23]. Traditionally, survival analysis relied on the semiparametric Cox Proportional Hazards (CPH) model [24]. However, CPH struggles with high-dimensional data, strict proportional hazard assumptions, and complex non-linear interactions [23], [25], [26]. Conversely, tree-based ensembles and boosting techniques handle non-linear patterns more effectively [27], [28]. Addressing data noise and missingness via advanced preprocessing is vital for model stability [29], [30], while compressive feature selection isolates a parsimonious core that preserves discriminative power [31]–[33]. To address these demands, this study presents a data engineering pipeline to predict 5-year BC survival in a cohort of 559 patients from Zona da Mata Mineira, Brazil [8]–[10]. Our analysis includes longitudinal hospital records covering baseline diagnostic parameters, anatomical staging, clinical treatment variables, and long-term follow-up status. Through this approach, we evaluate how feature engineering and selection isolate routine care predictors to enhance oncological decision-support models using Python's Scikit-survival library [34].

II. MATERIALS AND METHODS

The experimental pipeline follows a structured five-phase data engineering workflow (Figure 1): (i) cohort recruitment and filtering, (ii) stratified train-test partitioning, (iii) leakage-free preprocessing and imputation, (iv) benchmarking feature selection strategies, and (v) model training, validation, and uncertainty quantification. All routines were implemented in Python using Scikit-learn [35] and Scikit-survival [34] for right-censored time-to-event data.

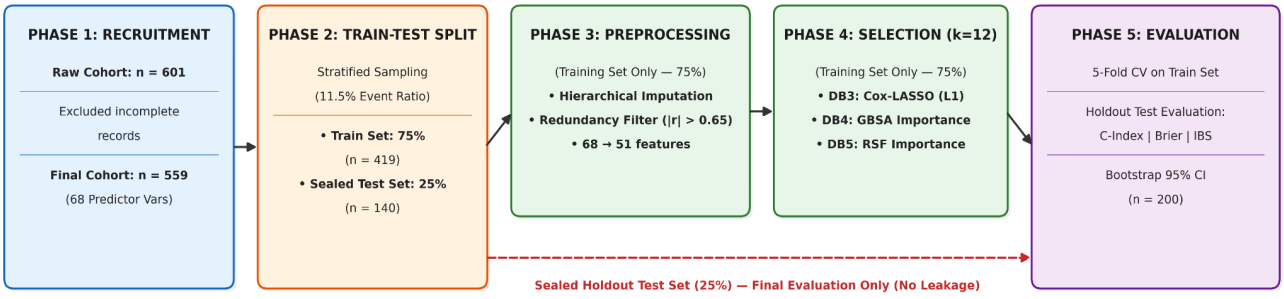


Figure 1. Methodological workflow of the 5-phase data engineering pipeline for 5-year female BC survival prediction, highlighting train-isolated preprocessing and the sealed holdout test set evaluation.

1. Study Population and Cohort Design

Clinical data derived from a retrospective hospital cohort of women with invasive BC (ICD-10: C50) in Zona da Mata Mineira, Brazil, were approved by the UFJF Ethics Committee (no. 5,533,296). Recruitment spanned three prior phases [8]–[10]: Phase 1 (2009–2010) extracted 601 records diagnosed in 2003–2005; Phase 2 (2010–2011) excluded cases without therapeutic follow-up; Phase 3 tracked vital status via SIM/DATASUS and telephone contact, establishing 563 patients. During the preprocessing phase of the present study, records with incomplete administrative or clinical data were excluded, yielding a final analytical cohort of 559 patients

[14], [31]. The feature matrix includes baseline diagnostic parameters (age, grade, molecular subtype, TNM) and longitudinal treatment trajectories (surgery, chemotherapy, radiotherapy) [8]–[10]. Time-to-event was the interval from diagnosis to BC death or 5-year right-censoring (1,825 days).

2. Leakage-Free Data Partitioning and Imputation

To eliminate data leakage, the dataset was partitioned at the onset into a 75% training set (n = 419) and a 25% holdout test set (n = 140) using stratified random sampling to preserve the 11.5% event-to-censoring ratio. The test set was frozen and sealed during all subsequent processing steps:

- **Baseline Raw Data (Database 1 - DB1):** Standardized time fields and removed attributes with missingness exceeding 40%, yielding 68 predictor variables.
- **Train-Isolated Imputation (Database 2.1 - DB2.1) & Filtering (Database 2 - DB2):** Missing values were imputed strictly within training boundaries using hierarchical clinical rules ($|r| > 0.65$) and statistical imputation (mode/median). Bivariate association analysis eliminated 17 redundant attributes ($|r| > 0.65$), reducing the matrix from 68 to 51 features.

3. Parsimonious Framework for Feature Selection

To identify clinically interpretable prognostic cores, selection algorithms were executed exclusively within the training partition:

- **Database 3 - DB3 (Cox-LASSO):** Selected via strict L_1-penalized shrinkage (l1_ratio = 1.0, alpha_min_ratio = 0.01).
- **Database 4 - DB4 (GBSA):** Tree-based Gradient Boosting ranked by Out-Of-Bag (OOB) permutation importance (15 repetitions).
- **Database 5 - DB5 (RSF):** Non-parametric Random Survival Forest (RSF) ranked by OOB permutation importance (15 repetitions).

Data-Driven Parsimony Criterion (k = 12): The cutoff was set by the sparsity boundary of Componentwise Gradient Boosting Survival Analysis (GBSA), which identified 12 active predictive attributes. Consequently, the *top 12* variables were standardized across DB3 through DB5 to enable a homogeneous comparison.

4. Predictive Models and Parameterization

Mathematical modeling integrated right-censorship [34] across linear, regularized, and ensemble paradigms. All models were initialized with a fixed seed (SEED = 42) and evaluated via 5-fold cross-validation strictly within the training partition (n = 419):

- **Regularized Cox Models:** Based on the classical CPH framework [24], regularized variants incorporate penalty terms to manage multicollinearity and stabilize coefficient estimation [36].
    - **Cox Proportional Hazards (CPH):** Configured with a light Ridge L_2 penalty ($\alpha = 10^{-4}$) to guarantee matrix inversion stability, using Breslow's baseline hazard estimator (ties = 'breslow', n_iter = 100) [36].
    - **Cox Elastic Net (CPH-EN):** Evaluated over a cross-validated path of 100 α values (alpha_min_ratio = 0.01) with an equal penalty balance (l1_ratio = 0.5) via coordinate descent (max_iter = 1000, fit_baseline_model = True) [36].
- **Ensemble Models:** Ensemble approaches capture non-linear relationships and complex interactions without relying on proportional hazards assumptions [37]–[42].
    - **Gradient Boosting Survival Analysis (GBSA):** Leverages Friedman's additive boosting framework [37], [38] adapted for Cox's partial likelihood loss function [39]. The model was built

using 250 decision trees (max_depth = 3) with regularization constraints (min_samples_split = 10, min_samples_leaf = 5) and a conservative learning_rate = 0.05.

- **Random Survival Forest (RSF):** Extends random forests to right-censored outcomes [40], constructing using 250 non-parametric survival trees (min_samples_split = 10, min_samples_leaf = 15, max_features = 'sqrt') guided by log-rank split criteria and Nelson-Aalen hazard estimation [40]–[42].

5. Model Evaluation and Uncertainty Quantification

Model performance was validated using *Scikit-survival* [34], across the 25% untouched holdout test set (n = 140) to evaluate both discriminative capacity and probability calibration:

- **Discriminative Capacity:** Longitudinal discrimination was measured using Harrell's and Uno's Concordance Index (C-index) [43]. The metric evaluates the application of Inverse Probability of Censoring Weighting (IPCW) to eliminate censoring bias (0.5 = random chance; 1.0 = perfect ordering).
- **Calibration and Prediction Error:** Calibration accuracy was assessed via the time-dependent Brier Score (BS) at the 5-year threshold (t = 1,825 days) and the Integrated Brier Score (IBS) evaluated across the continuous trajectory horizon ($t \in [0; 1{,}825]$ days) [44]. These metrics compute the IPCW mean-squared error between the actual status and the predicted survival probabilities, with values near zero indicating optimal global calibration.
- **Uncertainty Quantification:** Discriminative accuracy on training (C_{train}) was paired with nonparametric bootstrap resampling (nbootstraps = 200, SEED = 42) on the test partition (C_{test}) to calculate empirical 95% Confidence Intervals (95% CI) [CI_{lower}, CI_{upper}].

III. RESULTS AND DISCUSSION

The experimental evaluation benchmarked four survival modeling paradigms across five dataset configurations (k = 68, 51, 12) on the sealed n = 140 holdout test partition (20.0% event rate, 28 deaths in 5 years).

1. Cohort Characteristics and Feature Selection Dynamics

The curated cohort (n = 559) recorded 111 BC-related deaths over 5 years (79.75% right-censored rate), aligning with regional epidemiological indicators [8], [10]. Initial filtering (>40\% missingness) yielded 68 prognostic covariates (DB 1). Bivariate correlation mapping ($|r| > 0.65$) eliminated 17 redundant attributes, reducing the predictor space to 51 variables (DB 2) to prevent coefficient variance inflation [18], [36].

Feature selection algorithms compressed the operational space to distinct k = 12 attribute cores (Figure 2):

- **Cox-LASSO/Elastic Net Core (DB 3):** Exhibited 100% feature overlap (12/12) between both methods. The top 5 predictors were primary tumor size, post-surgical pathologically involved lymph nodes, semi-quantitative ER status (graded in crosses), distant metastasis (TNM M-category), and time to radiotherapy onset **(Figure 2A)**.
- **GBSA Permutation Core (DB 4):** The top 5 key drivers included treatment sequence modality (1-neoadjuvant, 2-adjuvant, 3-neoadjuvant and adjuvant, 4-palliative), tumor size, post-surgical pathologically involved lymph nodes, ER status (graded in crosses), and age at diagnosis **(Figure 2B)**.
- **RSF Permutation Core (DB 5):** The top 5 predictors were tumor size, immunohistochemical subtype (0-Luminal A, 1-Luminal B, 2-HER2+, 3-Triple-Negative, 4-Unknown), ER status (graded in crosses), post-surgical involved lymph nodes, and Ki-67 proliferation index **(Figure 2C).**

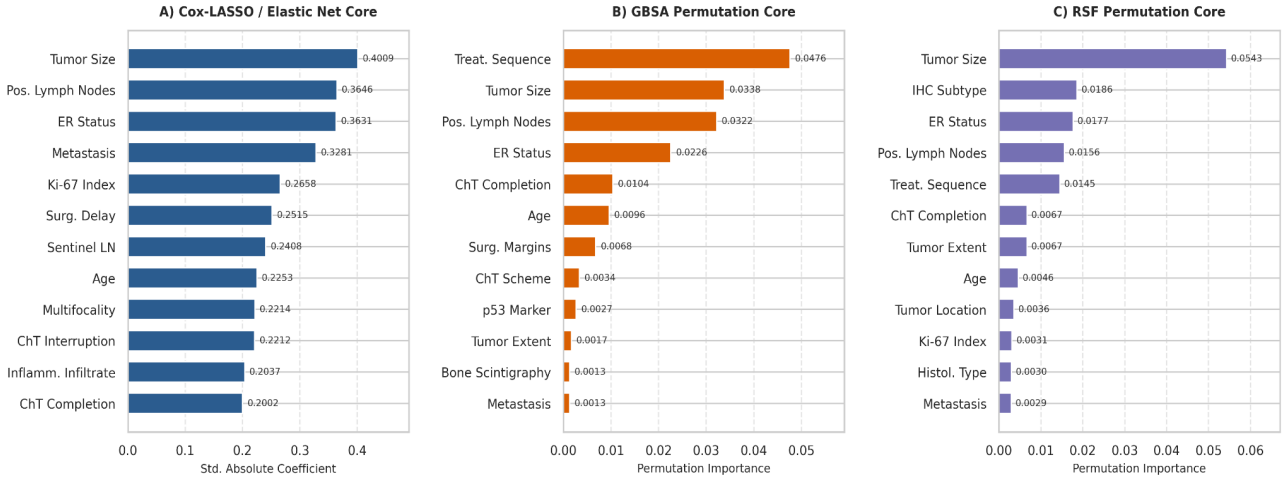


Figure 2. Relative importance ranking for the top k = 12 selected attributes across (A) Cox-LASSO, (B) GBSA, and (C) RSF feature selection paradigms (n = 140). Abbreviations: ER = Estrogen Receptor; ChT = Chemotherapy; LN = Lymph Node; Pos. = Positive; IHC = Immunohistochemical.

2. Comparative Analysis and Key Findings

The overall predictive performance and discrimination metrics across all dataset configurations and survival algorithms are summarized in Table 1. RSF achieved maximum overall discrimination on unrefined data (DB 1; k = 68, Ctest = 0.8879, 95% CI [0.8381, 0.9418]). On the other hand, standard CPH showed a moderate drop in discriminative ability on DB 1 (Ctest = 0.7647; BS = 0.1319), illustrating how high-dimensional clinical noise adversely affects traditional Cox regression [15], [24]. Feature filtering to k = 12 using GBSA permutation importance (DB 4) substantially improved CPH performance, raising its Ctest from 0.7647 to 0.8718 (95% CI [0.8219, 0.9195]) and reducing its 5-year BS to 0.0975 (IBS = 0.0823). Within this subset, GBSA achieved the best performance (Ctest = 0.8740, 95% CI [0.8206, 0.9261]). On the RSF permutation core (DB 5; k = 12), RSF demonstrated superior parsimonious generalization (Ctest = 0.8753, 95% CI [0.8214, 0.9268]), maintaining time-dependent Brier Scores well below the standard 0.25 clinical threshold (Table 1) [28], [44].

Table 1. Model Performance Across Feature Selection Strategies (n = 140 Holdout Test Set)

Paradigm	Model	*K*	C_{train}	C_{test}	95% CI Test	BS (5Y)	IBS
DB 1	CPH	68	0.8660	0.7647	[0.6800, 0.8588]	0.1319	0.1066
	CPH-EN		0.8436	0.8703	[0.8208, 0.9270]	**0.0901**	**0.0743**
	GBSA		0.9200	0.8550	[0.8025, 0.9148]	0.1061	0.0838
	RSF		0.8612	**0.8879**	[0.8381, 0.9418]	0.0993	0.0860
DB 2	CPH	51	0.8474	0.8093	[0.7181, 0.8957]	0.1074	**0.0826**
	CPH-EN		0.8047	0.8404	[0.7809, 0.9034]	**0.1012**	0.0839
	GBSA		0.9182	0.8574	[0.8030, 0.9147]	0.1085	0.0896
	RSF		0.8702	**0.8633**	[0.8066, 0.9167]	0.1159	0.0936
DB 3	CPH	12	0.8287	0.8204	[0.7371, 0.8982]	0.1033	0.0839
	CPH-EN		0.8086	0.8401	[0.7800, 0.9007]	0.1088	0.0907
	GBSA		0.8973	**0.8687**	[0.8157, 0.9251]	**0.0983**	**0.0811**

	RSF		0.8605	0.8586	[0.8055, 0.9083]	0.1124	0.0906
DB 4	CPH	12	0.8029	0.8718	[0.8219, 0.9195]	**0.0975**	**0.0823**
	CPH-EN		0.7954	0.8515	[0.7949, 0.9121]	0.1005	0.0847
	GBSA		0.9015	**0.8740**	[0.8206, 0.9261]	0.0997	0.0867
	RSF		0.8523	0.8694	[0.8156, 0.9302]	0.1033	0.0878
DB 5	CPH	12	0.8104	0.8483	[0.7904, 0.9058]	0.1018	0.0839
	CPH-EN		0.8011	0.8404	[0.7845, 0.8924]	0.1022	0.0858
	GBSA		0.9054	0.8555	[0.8044, 0.9153]	0.1043	0.0813
	RSF		0.8519	**0.8753**	[0.8214, 0.9268]	0.1017	0.0865

Note: DB 1: Unfiltered Raw Data; DB 2: Correlation Filtered; DB 3: Cox-LASSO / Elastic Net Core; DB 4: GBSA Permutation Core; DB 5: RSF Permutation Core. k = Active Features; C_{train} and C_{test} = Uno's Concordance Index on train (n=419) and test (n=140) sets; CI TEST = Confidence Interval (200 bootstrap resamples); BS (5Y) = Brier Score at t = 1,825 days; IBS = Integrated Brier Score over 5 years. Models: CPH = Cox Proportional Hazards; CPH-EN = CPH with Elastic Net Penalty; GBSA = Gradient Boosting Survival Analysis; RSF = Random Survival Forest. Bold indicates optimal metrics per dataset.

Reducing the feature space from k = 68 to k = 12 demonstrated that a compact core of variables retains the essential clinical signal for 5-year survival prediction. While RSF achieved maximum discrimination on raw data (Ctest = 0.8879 on DB 1), feature selection enabled highly accurate parsimonious models, such as GBSA on DB 4 (Ctest = 0.8740) and RSF itself on DB 5 (Ctest = 0.8753), while restoring the predictive capacity of classical CPH (Ctest = 0.8718), all while maintaining time-dependent calibration errors well below clinical validation thresholds [28], [44].

3. Discussion and Clinical Interpretation

The transition from classical biostatistics to ML survival algorithms on RWD is pivotal for precision oncology [14], [26]. The proportional hazards assumption inherent to standard CPH regression explains its baseline limitations on unrefined data (DB 1; Table 1), as such assumptions are frequently violated in real-world registries characterized by heterogeneous treatment courses and time-varying covariates [23]. In contrast, non-parametric tree ensembles and boosting methods dynamically capture complex non-linear clinical interactions [28], [44]. Our top-performing models on compact subsets (Ctest = 0.8753 for RSF on DB 5; Ctest = 0.8740 for GBSA on DB 4) consistently outperform benchmark models reported in recent literature. For instance, Kang et al. achieved a maximum C-index of 81.23\% (95% CI [80.34–82.12]) for RSF and 81.73\% (95% CI [81.73–81.73]) for Cox-Elastic Net in their older cohort, while dropping to 74.91\% (95% CI [72.47–77.36]) and 72.61\% (95% CI [72.61–72.61]) in younger patients. Similarly, Cai et al. [20] evaluated performance across varying subsets of variables (k = 8 to 16), reporting test-set C-indices ranging from 0.785 to 0.787 for RSF, 0.796 to 0.803 for CoxPH, and 0.767 to 0.794 for DeepSurv. Notably, DeepSurv exhibited clear overfitting in their evaluation, peaking at a training C-index of 0.845 but dropping to 0.769 on the test set (k = 14). In contrast, our pipeline maintained high discriminative performance (0.8753) and stability on the held-out test set using a parsimonious core of 12 clinical attributes, while deliberately avoiding deep learning architectures that lack clinical interpretability [18], [20]. Model stability and calibration were robust. In contrast to studies showing high metric values during training (e.g., Li et al. [33], where RSF reached a training C-index of 0.920 and a 3-year AUC of 0.947), our primary models maintained generalizability on the sealed holdout test partition (n = 140), with Ctest $\geq$ 0.8694. Time-dependent calibration errors remained low (BS $\leq$ 0.1033; IBS $\leq$ 0.0878 over 5 years; Table 1). This performance aligns with our previous baseline (Carvalho et al. [27]), which reported IBS values of 0.0547 for RSF and 0.0437 for GBSA with 70 predictors. It compares favorably with recent reports such as [30], in which the best model for overall survival (GB) achieved a C-index of 0.68 (95% CI [0.57–0.79]) and an IBS of

0.17. Feature selection grounded in clinical consensus mitigated optimism bias while preserving temporal calibration within clinical validation thresholds.

Importantly, automated feature selection converged with core biological domains (Figure 2) [7], [10]:

- **Anatomical and Pathological Staging:** Tumor size, positive lymph nodes, tumor extent, and distant metastasis reflect disease burden within the standard TNM framework (**Figure 2A-C**) [8], [9].
- **Biomarkers and Molecular Subtypes:** ER status, Ki-67 proliferation index, p53 marker, and immunohistochemical subtypes capture endocrine responsiveness and biological aggressiveness (**Figure 2A, 2C**) [6], [7].
- **Care Trajectories:** Treatment sequence, surgical delays, chemotherapy schemes, and cycle completion highlight the direct clinical impact of healthcare delivery timelines and therapeutic continuity on 5-year survival outcomes (**Figures 2A and 2B**) [6], [33].

IV. CONCLUSION

This study validates the use of supervised ML to predict 5-year overall survival in an invasive BC cohort. Non-parametric ensemble algorithms, led by RSF on the unrefined dataset (DB 1: *C*-index = 0.8879) and on the reduced permutation core (DB 5: *C*-index = 0.8753), demonstrated superior discrimination and robust temporal calibration (BS ≤ 0.1017, IBS ≤ 0.0865), outperforming standard CPH regression. Crucially, compressing the operational space from $k = 68$ to $k = 12$ clinical attributes preserved 98.2% of peak performance while recovering CPH calibration (BS = 0.0975 on DB 4). This confirms that accurate, interpretable prognostic tools can be built using routine clinical parameters, offering a scalable decision-support alternative for resource-constrained healthcare systems [6], [31]. Limitations include the retrospective single-center design and the lack of external validation. However, this risk is mitigated by the study site's representativeness: a major tertiary oncology center serving both the public (SUS) and private sectors that closely reflects the demographics of the broader Brazilian population. Additionally, historical registries underrepresent recent therapeutic advances, which may shift baseline survival [28]. Future work will focus on multi-institutional validation, dynamic imputation techniques, and local Explainable AI frameworks (SHAP and LIME) to enhance point-of-care interpretability [14].

ACKNOWLEDGEMENT

This study was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, through a Doctoral Fellowship and the Institutional Sandwich Doctorate Program Abroad (PDSE) Grant. The authors thank the Faculty of Engineering of the University of Porto (FEUP) and the Graduate Program in Computational Modeling at the Federal University of Juiz de Fora (PGMC/UFJF) for their academic support; Prof. Dr. Maximiliano Ribeiro Guerra and Dr. Jane Rocha Duarte Cintra for their valuable collaboration and for providing the clinical database; and Hospital 9 de Julho and Instituto Oncológico for their institutional support.

REFERENCES

[1]. F. Bray, M. Laversanne, E. Weiderpass, and J. Ferlay, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2]. Instituto Nacional de Câncer (INCA), "Estimativa 2026: Incidência de Câncer no Brasil," Rio de Janeiro: INCA, 2026.

[3]. A. Miranda-Filho et al., "Global patterns and trends in breast cancer incidence and mortality," *The Breast*, vol. 71, p. 103562, 2025.

[4]. International Agency for Research on Cancer (IARC), "Cancer Today," World Health Organization, 2026. [Online]. Available: https://gco.iarc.fr/today

[5]. Brasil, Ministério da Saúde, "Sistema de Informações sobre Mortalidade (SIM/DATASUS) - Estatísticas de Mortalidade por Câncer de Mama", DATASUS, 2025.

[6]. Brazil, *Clinical Protocols and Therapeutic Guidelines in Oncology*, M. I. P. Gadelha, S. J. Martins, and P. D. Picon, Eds. Brasília, DF: Ministério da Saúde, 2014.

[7]. C. H. E. Barrios et al., "Breast: Staging," in *Oncological Treatment Guidelines Recommended by the Brazilian Society of Clinical Oncology (SBOC)*, 2022.

[8]. A. S. S. Ferreira, J. R. D. Cintra, V. A. Fayer, M. C. Nogueira et al., "Breast cancer survival and the health system in Brazil: an analysis of public and private healthcare," *Front. Oncol.*, vol. 13, p. 927748, May 2023.

[9]. J. R. D. Cintra et al., "Perfil imuno-histoquímico e variáveis clinicopatológicas no câncer de mama," *Revista da Associação Médica Brasileira*, vol. 58, no. 2, pp. 178–187, 2012.

[10]. V. A. Fayer, M. R. Guerra, J. R. D. Cintra, and M. T. Bustamante-Teixeira, "Sobrevida de dez anos e fatores prognósticos para o câncer de mama na região Sudeste do Brasil," *Rev. Bras. Epidemiol.*, vol. 19, no. 4, pp. 766–778, oct. 2016.

[11]. M. T. Bustamante-Teixeira, E. Faerstein, and M. R. Latorre, "Técnicas de análise de sobrevida," *Cad. Saúde Pública*, vol. 18, no. 3, pp. 579–594, May 2002.

[12]. C. Allemani et al., "Global surveillance of trends in cancer survival 2000–2014 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries," *The Lancet*, vol. 391, no 10125, pp. 1023–1075, 2018.

[13]. F. Kitsios, M. Kamariotou, and M. Syngelakis, "Recent advances of artificial intelligence in healthcare: a systematic literature review," *Appl. Sci.*, vol. 13, no. 13, p. 7479, Jun. 2023, doi: 10.3390/app13137479.

[14]. J. Li et al., "Predicting breast cancer 5-year survival using machine learning: a systematic review," *PLoS ONE*, vol. 16, no. 4, p. e0250370, 2021.

[15]. N. Min, Y. Wei, Y. Zheng, and X. Li, "Advancement of prognostic models in breast cancer: a narrative review," *Gland Surgery*, vol. 10, no. 9, pp. 2815–2828, 2021.

[16]. G. M. M. Paixão et al., "Machine learning na medicina: revisão e aplicabilidade," *Arquivos Brasileiros de Cardiologia*, vol. 118, pp. 95–102, 2022.

[17]. D. S. Carvalho, T. S. Nogueira, and P. V. Z. C. Goliatt, "Aplicação do random survival forest na análise da sobrevida para câncer da mama," *Journal of Health Informatics*, vol. 15, no. Especial, 2023, doi: 10.59681/2175-4411.v15.iEspecial.2023.1113.

[18]. L. Cattelani and V. Fortino, "Dual-stage optimizer for systematic overestimation adjustment applied to multi-objective genetic algorithms for biomarker selection," *Brief. Bioinform.*, vol. 26, no. 1, p. bbae674, Jan. 2025.

[19]. E. Curioso et al., "Dimensionality reduction and feature consensus in clinical oncology survival data," *Bioinformatics*, vol. 39, no. 5, p. btad302, 2023.

[20]. W. Cai, Y. Qi, L. Zheng et al., "Comparison of random survival forest-based overall survival with deep learning and Cox proportional hazard models in HER-2-positive HR-negative breast cancer," *Cancer Rep.*, vol. 8, no. 7, p. e70262, Jul. 2025, doi: 10.1002/cnr2.70262.

[21]. T. Jansen, G. Geleijnse, M. Van Maaren, M. P. Hendriks, A. Ten Teije, and A. Moncada-Torres, "Machine learning explainability in breast cancer survival," in *Digital Personalized Health and Medicine*, Amsterdam: IOS Press, 2020, pp. 307–311, doi: 10.3233/SHTI200172.

[22]. M. Krzyziński et al., "SurvSHAP(t): time-dependent explanations of machine learning survival models," *Knowledge-Based Systems*, vol. 262, p. 110234, 2023, doi: 10.1016/j.knosys.2022.110234.

[23]. A. Moncada-Torres et al., "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," *Scientific Reports*, vol. 11, no. 1, p. 14918, 2021.

[24]. D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[25]. A. Fanizzi, D. Pomarico, A. Rizzo, S. Bove, M. C. Comes, V. Didonna et al., "Machine learning survival models trained on clinical data to identify high-risk patients with hormone-responsive HER2-negative breast cancer," *Scientific Reports*, vol. 13, no. 1, p. 8575, 2023, doi: 10.1038/s41598-023-35619-7.

[26]. P. Liu et al., "Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 1, pp. 148–160, 2020.

[27]. D. S. Carvalho, P. V. Z. C. Goliatt, and L. Goliatt, "Comparative analysis of machine learning models for breast cancer patients' survival prediction," in *Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA 2024)*, Cham: Springer, 2024, pp. 172–181.

[28]. J. Xiao et al., "The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study," *JMIR Medical Informatics*, vol. 10, no. 2, p. e33440, 2022.

[29]. H. Y. J. Kang, M. Ko, and K. S. Ryu, "Prediction model for survival of younger patients with breast cancer using the breast cancer public staging database," *Sci. Rep.*, vol. 14, p. 25723, 2024.

[30]. L. Mastrantoni, G. Garufi, G. Giordano et al., "Comparison of machine learning and deep learning models for survival prediction in early-stage hormone receptor-positive/HER2-negative breast cancer receiving neoadjuvant chemotherapy," *ESMO Real World Data Digit. Oncol.*, vol. 10, p. 100184, 2025.

[31]. H. I. Okagbue, P. I. Adamu, P. E. Oguntunde, E. C. Obasi, and O. A. Odetunmibi, "Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis, and location of cancer," *Health and Technology*, vol. 11, no. 1, pp. 1–7, 2021, doi: 10.1007/s12553-020-00489-x.

[32]. Z. Sedighi-Maman and A. Mondello, "A two-stage modeling approach for breast cancer survivability prediction," *International Journal of Medical Informatics*, vol. 149, p. 104438, 2021.

[33]. L. W. Li, X. Liu, M. L. Shen, M. J. Zhao, and H. Liu, "Development and validation of a random survival forest model for predicting long-term survival of early-stage young breast cancer patients based on the SEER database and an external validation cohort," *Am. J. Cancer Res.*, vol. 14, no. 4, pp. 1609–1621, Apr. 2024.

[34]. S. Pölsterl, "Scikit-survival: a library for time-to-event analysis built on top of scikit-learn," *Journal of Machine Learning Research*, vol. 21, no. 212, pp. 1–6, 2020.

[35]. F. Pedregosa et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36]. N. Simon et al., "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–22, 2011.

[37]. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[38]. J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002, doi: 10.1016/S0167-9473(01)00065-2.

[39]. T. Hothorn et al., "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2006.

[40]. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41]. H. Ishwaran and U. B. Kogalur, "Random survival forests for R," *R News*, vol. 7, no. 2, pp. 17–31, Oct. 2007.

[42]. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, 2008, doi: 10.1214/08-AOAS169.

[43]. Y. Uno et al., "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no 10, pp. 1105–1117, 2011.

[44]. M. Graf et al., "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine*, vol. 18, no. 17–18, pp. 2529–2545, 1999.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Precise Segmentation of Dental Implants with YOLO-Based Deep Learning Models

Hüseyin POLAT
Department of Computer Engineering
Faculty of Technology, Gazi University
Ankara, Turkey
polath@gazi.edu.tr
ORCID:0000-0003-4128-2625

Hakan ÜNAL
Graduate School of Natural and Applied Sciences
Gazi University
Ankara, Turkey
hakan.unal1@gazi.edu.tr
ORCID:0000-0003-3961-9663

Emin UYSAL
Pi Health Technologies
Ankara, Turkey
emin@pixr.store
ORCID:0009-0001-8600-6459

Elif Meltem ASLAN
Hamidiye Faculty of Dentistry
University of Health Sciences
İstanbul, Turkey
aslan.meltem5@gmail.com
ORCID:0000-0002-1737-9585

Abstract
Dental implants are a widely used treatment modality for replacing missing teeth. The implant placement procedure requires accurate positioning, which should be supported by precise, detailed dental imaging analysis. This study aims to perform automatic dental implant segmentation using YOLO (You Only Look Once)-based deep learning models and to compare their performance. Compared with manual analysis, automatic segmentation enables faster, more consistent assessment while reducing clinicians' workload and potentially allowing care to be provided to more patients. In this study, YOLOv8, YOLOv9, YOLOv11, and YOLOv12 were trained and tested on dental images. Model performance was evaluated using accuracy, precision, recall, and processing-time metrics. YOLOv8 demonstrated a balanced performance, achieving a mask mAP@0.50:0.95 of 96.7% with an inference time of 10.309 ms per image. YOLOv9 yielded the highest mask segmentation performance, reaching 97.0% mAP@0.50:0.95, although its inference time was longer at 25.072 ms per image. YOLOv11 achieved a comparable mask mAP@0.50:0.95 of 96.9% with an inference time of 24.825 ms per image. YOLOv12 attained a mask mAP@0.50:0.95 of 93.9%, 99.9% precision, 100% recall, an inference time of 10.831 ms per image, and an approximate throughput of 92.3 FPS. This comparative analysis is expected to help identify the most effective and reliable deep learning model for dental implant segmentation.

Keywords: YOLO, Dental Implant, Deep Learning, Artificial Intelligence, Segmentation

I. INTRODUCTION

Dental implants constitute a widely adopted and effective therapeutic approach for the rehabilitation of missing teeth in contemporary dentistry. Implants, typically manufactured from biocompatible materials such as titanium, are placed into the jawbone and subsequently restored with a prosthesis, thereby providing both functional and esthetic outcomes. Nevertheless, because the success of implant therapy depends on accurate implant positioning and appropriate osseointegration with the surrounding bone, precise and detailed analysis of dental images at the initial stage of treatment planning is of critical importance [1]. In dental implant planning, although panoramic radiography provides a two-dimensional overview for general assessment, it may be insufficient for precise measurement of parameters such as bone density, width, and height due to magnification, geometric distortion, and inherent resolution limitations [1]. Therefore, Cone-Beam Computed Tomography (CBCT) is often preferred for more comprehensive anatomical evaluation; by enabling three-dimensional visualization of the jawbone, CBCT facilitates more accurate assessment of bone quality and

critical structures such as the mandibular canal and the maxillary sinus, thereby supporting the optimization of surgical planning [1]. Nevertheless, manual analysis of CBCT data is time-consuming due to the large volume of information; moreover, it may lead to inconsistencies arising from subjective variability influenced by clinician experience, attention, and fatigue, as well as human error. Given these constraints, deep learning (DL)–based approaches offer substantial advantages by enabling faster analysis of CBCT images, enhancing standardization, and yielding reproducible outcomes. Segmentation, one of the fundamental applications of DL, enables the automatic delineation of teeth, jawbone, and related anatomical structures in dental images, thereby facilitating the identification of suitable regions for implant placement and contributing to a more objective surgical planning process [2]. In recent years, automatic segmentation applications have provided clinical decision support in areas such as implant site localization, prediction of grafting requirements, and reduction of the risk of nerve injury by producing faster and more consistent results compared with manual assessment [3].

In CBCT-based analyses, the YOLO (You Only Look Once) family offers substantial potential for high performance and real-time efficiency in object detection and segmentation tasks owing to its single-stage architecture, making it a suitable candidate for precision-demanding applications such as dental implant segmentation. In this study, YOLOv8, YOLOv9, YOLOv11, and YOLOv12 were compared for dental implant segmentation, and their performance was evaluated using metrics including precision, recall, F1-score, mean average precision, and processing time. The findings reveal the respective strengths and limitations of the different YOLO versions and thereby help identify the most effective and reliable model for implant segmentation. Furthermore, deep learning-based segmentation is clinically valuable because of its potential to reduce clinicians' workload and provide safer decision support throughout surgical workflows.

II. LITERATURE REVIEW

[4] - For breast cancer mass segmentation in digital mammography, the YOLO-LOGO model was proposed by integrating YOLOv5 with Transformer-based Local–Global (LOGO) components. Evaluated on the CBIS-DDSM and INBreast datasets, the approach was reported to achieve superior detection/segmentation performance compared with prior methods while also reducing computational requirements. [5] - YOLOv2 and YOLOv3 were compared for object detection in aerial imagery, with analyses conducted on classes such as vehicles, aircraft, harbors, and ships using the DOTA dataset. The results indicated that YOLOv2 performed better on large objects, whereas YOLOv3 performed better on small objects. In addition, YOLOv3 demonstrated an advantage in inference speed (~2.5 s on average); however, for certain classes, YOLOv2 maintained higher accuracy. [6] - INSTA-YOLO was developed for real-time instance segmentation and operates in a single stage, in contrast to conventional two-stage approaches. By using a contour-based 2D point representation instead of pixel-wise masks, the method aimed to improve computational efficiency; across the Carvana, Cityscapes, and Airbus datasets, it was reported to achieve competitive accuracy while providing approximately a two-fold increase in speed. [7] - For segmentation in prostate cancer histopathology images, Mamba-, SAM-, and YOLO-based models were compared and evaluated on the Gleason 2019 and SICAPv2 datasets. The results indicate that the Mamba approach employing the H-vmunet architecture achieved superior performance in terms of the Dice coefficient. [8] - For the automated assessment of fish freshness, YOLO and Mask R-CNN were compared for segmenting fish head and tail regions. Using the Fish-gres dataset, YOLO was reported to achieve higher precision (98.96%) and recall (80.93%), yielding superior mAP values. [9] - The CNN family and the YOLO family were discussed in terms of architectural framework, computational speed, and object recognition effectiveness. It was emphasized that Faster R-CNN achieves strong detection performance, and that combining an RPN with Faster R-CNN is more advanced than the Selective Search + Faster R-CNN approach. In addition, YOLO was described as a CNN-based method capable of detecting multiple targets with a certain level of accuracy through end-to-end training and inference.

[10] - Two representative algorithm families for object detection, namely CNN-based and YOLO-based approaches, were presented and discussed along the accuracy–speed–cost trade-off. In particular, YOLOv3 has been reported to strike a balance between inference speed and detection accuracy.

[11] - A hybrid YOLO + GrabCut approach was proposed for skin cancer lesion segmentation; after localizing the lesion with YOLO, segmentation was completed using GrabCut. Evaluated on the ISBI 2017 and PH2 datasets, the method was reported to yield competitive results with approximately 90% precision. [12] - For real-time detection and classification of apple surface defects, a BiSeNet V2 + YOLOv4 framework was employed: BiSeNet V2 segmented the defective regions. At the same time, YOLOv4 served as a verification module to reduce the error rate. The results were reported as 92.42% accuracy and 94.31% F1-score.

[13] - For brain tumor segmentation, Inception-V3 and YOLO were integrated; homomorphic wavelet filtering was applied as a preprocessing step, and tumor localization was performed using a YOLOv2–InceptionV3 framework. On the BRATS 2018/2019/2020 datasets, IoU values of 97%, 98%, and 100% were reported, respectively.

[14] - The evolution of YOLO and its performance in real-time detection/segmentation were examined; YOLO versions from v1 to v8 were compared in terms of speed and accuracy on the COCO dataset. It was noted that each successive release introduced optimizations that led to performance improvements.

[15] - A cloud-based CNN was developed for the automatic segmentation of dental implants and associated prosthetic crowns in CBCT images. Training/validation/testing were conducted using 280 CBCT scans; the automated segmentation (AS) method was found to be approximately 60× faster than the semi-automatic approach (SS) (<30 s). Moreover, high segmentation accuracy for implants (DSC: 0.92 ± 0.02; 0.91 ± 0.03) and low error (at the RMSE level) were reported.

[16] - A hierarchical, multi-stage method was developed for fully automatic 3D individual tooth segmentation and identification from dental CBCT. To reduce complexity, panoramic-derived representations were generated for the maxilla and mandible; subsequently, 2D detection/segmentation was performed, followed by 3D segmentation using a coarse-to-fine ROI strategy. Reported performance included an F1-score of 93.35% for tooth detection and a Dice score of 94.79% for 3D tooth segmentation.

[17] - A CNN-based method was proposed to automate tooth segmentation required for the reconstruction of 3D dental models. The approach employed multi-stage training using different sub-volumes; to enhance compatibility across CBCT scanners, histogram-based preprocessing and a posterior probability function were applied. It was reported that performance was further improved by using spatial dropout and dense convolutional blocks.

[18] - In a systematic review evaluating AI applications in dental implant planning, it was emphasized that the automatic detection/segmentation of anatomical structures improves both accuracy and efficiency. Use cases such as automated identification of bone density, the maxillary sinus, neural structures, and teeth were discussed; however, the need for high-quality training data and the lack of protocol standardization were highlighted as major limitations.

[19] - A ResUNet-based model was developed for tooth segmentation and automated implant planning from panoramic and tangential images. PCA and linear regression were employed to detect missing teeth and recommend the implant axis; segmentation performance was 97.85% accuracy (IoU: 84.83%), and the mean implant site deviation was 0.36 mm.

[20] - A U-Net–based model was trained for the segmentation of teeth, restorations, implants, fillings, caries, root canals, and residual roots in OPG images. The reported performance included DSC values of 85% for teeth, 88% for caries, 87% for restorations, 93% for bridge restorations, and 94% for implants; the potential clinical contribution in terms of speed and accuracy was highlighted.

[21] - For implant planning in edentulous mandibles, a U-Net/MONAI-based automated segmentation approach was developed to delineate bone volume in CBCT images. Training and testing were performed using 43 CBCT scans; DSC values of 0.89 for the training set and 0.78 for the test set were reported. Higher performance was observed in unilateral cases, whereas limitations were noted for bilateral cases.

[22] – For segmenting implants and peri-implant crestal bone in radiographic images, a pipeline was proposed in which noise reduction, morphological operations, and histogram thresholding were followed by final segmentation using an Active Shape Model (ASM). Compared with manual measurements, an ICC of 0.75 and 95% agreement in the Bland–Altman analysis were reported, with a mean difference of 0.049 mm.

[23] – For implant (titanium rod) segmentation in CBCT, a three-stage automated method was developed, comprising (i) Tmin pre-thresholding, (ii) edge enhancement using the Sobel operator, and (iii) thresholding with a fixed Tedge of 7.5%. Validated on a phantom across 17 CBCT device protocols with sub-millimeter accuracy, the approach was reported to detect all rods except on one device correctly.

III. MATERIALS AND METHODS

This study aims to compare the performance of the YOLO-based deep learning models YOLOv8, YOLOv9, YOLOv11, and YOLOv12 for dental implant segmentation. These approaches are used for multi-object detection and instance segmentation and are optimized differently with respect to the accuracy–speed trade-off because of differences in their core architectures. The study was conducted using an ethically approved dental CBCT dataset. The images were curated in Roboflow to meet YOLO training requirements, and the corresponding TXT and YAML annotation files were generated. During preprocessing, Auto-Orient was applied, and all images were resized to 640 × 640 pixels. Training and evaluation were performed in a CUDA-enabled PyTorch environment. Performance assessment was based on bounding-box and mask precision, recall, F1-score, mAP@0.50, mAP@0.50:0.95, and preprocessing, inference, and postprocessing times.

1. Data Set

The dataset used in this study comprises dental images obtained from Medipol University Faculty of Dentistry and Dental Hospital (Ankara, Türkiye). Using the Roboflow platform, the images were partitioned into training (70%), validation (10%), and test (20%) sets; the training set was used to learn implant segmentation, and the final performance evaluation was conducted on the test set. The dataset includes 305 images from 9 individuals (7 females, 2 males). The training subset consists of 214 images and covers all 9 individuals (7 females, 2 males), while the test subset contains 91 images from 8 individuals (6 females, 2 males) and was used to assess model outputs objectively.

2. YOLOv8 Segmentation Model

YOLOv8 (You Only Look Once, v8) is an advanced deep learning architecture developed for object detection and segmentation tasks, optimized for both speed and accuracy [24]. In this study, the model was implemented in PyTorch and trained specifically for dental implant segmentation, with performance evaluated using metrics such as accuracy and inference time. Within the architectural context summarized in Figure 1, the main components of YOLOv8 can be highlighted as follows: (i) an anchor-free detection paradigm that directly predicts object centers, (ii) updates to the Backbone/Neck design, including replacing the initial 6×6 convolution with a 3×3 convolution and adopting the C2f module instead of C3 (where Bottleneck outputs are aggregated in C2f), and (iii) the management/disabled setting of the mosaic augmentation mechanism, which is part of the online augmentation strategies used during training [25]. Moreover, YOLOv8's strong performance on benchmark datasets such as COCO and Roboflow 100 (e.g., 50.2% mAP on COCO for

YOLOv8m), its developer-friendly CLI/Python package structure, and extensive community support are key factors explaining its broad adoption in practical applications [25].

3. YOLOv9 Segmentation Model

YOLOv9 is a deep learning model optimized for object detection and segmentation tasks, and, in this study, it was adapted for dental implant segmentation [26]. Computational processes were accelerated using a CUDA-enabled GPU; to enhance data-adaptive learning during training, hyperparameters were set to batch size = 8, IoU threshold = 0.7, and confidence threshold = 0.5. Upon completion of training, YOLOv9 demonstrated high accuracy in implant segmentation. Among the principal architectural contributions that distinguish YOLOv9 from earlier versions are Programmable Gradient Information (PGI), which aims to reduce information loss in deep networks by promoting reliable gradient formation and improving training efficiency, and the Generalized Efficient Layer Aggregation Network (GELAN), which maximizes parameter utilization and computational efficiency. In addition, it has been reported that multi-level features are effectively fused within the Neck component via RepNCSPELAN4, which integrates CSPNet- and ELAN-like paradigms [27]. Collectively, these characteristics help explain why YOLOv9 establishes a competitive standard in real-time applications in the accuracy–speed–parameter efficiency trade-off and why it demonstrates strong performance on complex tasks such as dental implant segmentation [27].

4. YOLOv11 Segmentation Model

YOLOv11 is a recent YOLO variant used for dental implant segmentation in this study [28]. Training used a batch size of 16, with AMP (automatic mixed precision) disabled and the remaining hyperparameters kept similar to those of YOLOv9. Architecturally, YOLOv11 incorporates C3K2 blocks in the Backbone/Neck, SPPF (Spatial Pyramid Pooling Fast) for multi-scale contextual representation, and C2PSA (Channel and Spatial Attention) in the Head. These components improve parameter efficiency and small-object detection while maintaining a favorable speed–accuracy balance [29], supporting effective segmentation of complex dental images [29].

5. YOLOv12 Segmentation Model

YOLOv12 is an attention-centric YOLO variant that integrates attention mechanisms into the real-time object detection framework. The model employs Area Attention to capture broad spatial dependencies at reduced computational cost and a Residual Efficient Layer Aggregation Network (R-ELAN) to improve feature aggregation efficiency [30]. In this study, the YOLOv12n-seg configuration was adapted to a single-class dental implant segmentation task. The fused model comprised 172 layers, 2,809,339 parameters, and a computational complexity of 9.6 GFLOPs. Final testing was performed using Ultralytics 8.4.104, Python 3.11.9, and PyTorch 2.11.0+cu128 on an NVIDIA GeForce RTX 3060 12-GB GPU with CUDA 12.8 and cuDNN 9.1.9. The best-performing trained weights were evaluated on the 61-image test set, and bounding-box, segmentation-mask, and runtime metrics were calculated.

IV. MODEL PERFORMANCE EVALUATION METRICS AND COMPARATIVE ANALYSIS

In this study, YOLOv8, YOLOv9, YOLOv11, and YOLOv12, trained for dental implant segmentation, were evaluated on an independent test set for object detection and segmentation performance. Segmentation quality, localization sensitivity, and runtime profiles were considered jointly. The evaluation metrics included mAP@0.50 (Boxes/Masks), which represents mean average precision at an IoU threshold of 0.50; mAP@0.50:0.95 (Boxes/Masks), which reflects generalized performance over IoU thresholds ranging from 0.50 to 0.95; precision, which indicates the capacity to limit false-positive predictions; recall, which represents the capacity to avoid missed implants; and the F1-score, which summarizes the precision–recall balance. Runtime performance was assessed using preprocessing, inference, and postprocessing times. The resulting metrics are presented in Table 1.

Table 1. Performance metrics and durations of the models

Metric	**Models**			
	YOLOv8	***YOLOv9***	***YOLOv11***	***YOLOv12***
mAP@0.5 (Average Sensitivity at 0.5 IoU Threshold) (Boxes)	0.995	0.995	0.995	0.995
mAP@0.5:0.95 (Average Sensitivity in the Range of IoU 0.5–0.95) (Boxes)	0.963	0.961	0.945	0.931
Precision (Boxes)	0.999	0.998	0.999	0.999
Recall (Boxes)	1.000	1.000	1.000	1.000
F1-Score (Boxes)	1.000	0.999	1.000	1.000
mAP@0.5 (Masks)	0.995	0.995	0.995	0.995
mAP@0.5:0.95 (Masks)	0.967	0.970	0.969	0.939
Precision (Masks)	0.999	0.998	0.999	0.999
Recall (Masks)	1.000	1.000	1.000	1.000
F1-Score (Masks)	1.000	0.999	1.000	1.000
Preprocessing Time (ms) (test)	2.13 ms	2.10 ms	4.87 ms	6.39 ms
Inference Time (ms) (test)	10.31 ms	25.07 ms	24.82 ms	10.83 ms
Post-processing Time (ms) (test)	3.51 ms	3.05 ms	4.29 ms	3.70 ms

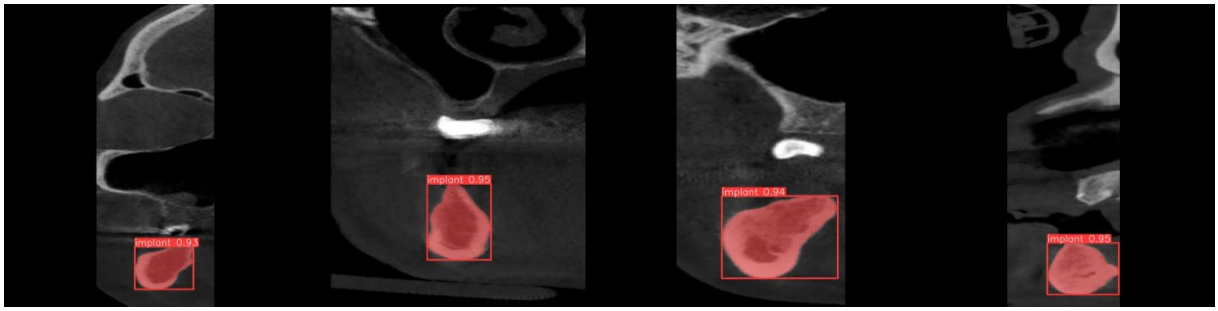


Figure 1. Sample dental implant segmentation results.

Figure 1 presents the results of the dental implant segmentation study. The figure illustrates the detection and masking of implants in different dental tomography slices using a YOLO-based segmentation model. The red boxes denote the bounding boxes of the detected implants, whereas the red overlay mask represents the segmented implant regions. The numerical values displayed on the boxes (e.g., 0.93) indicate the confidence scores corresponding to the model's predictions. Overall, these outputs demonstrate that the model achieves successful performance in both detection and segmentation.

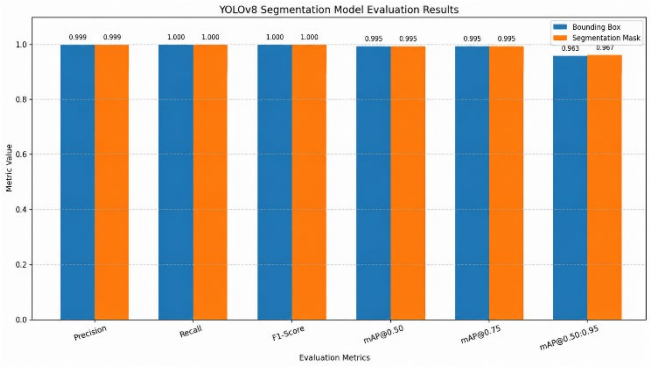


Figure 2. YOLOv8 segmentation results.

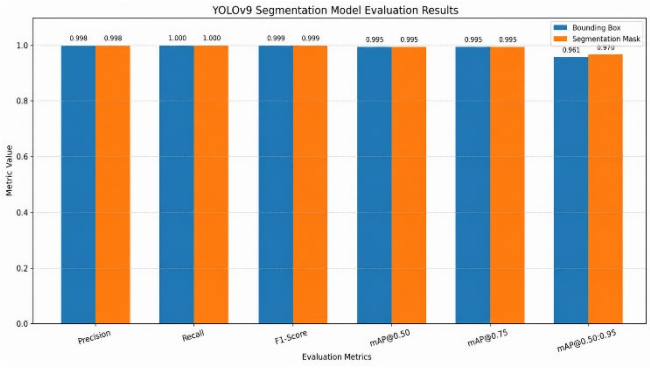


Figure 3. YOLOv9 segmentation results.

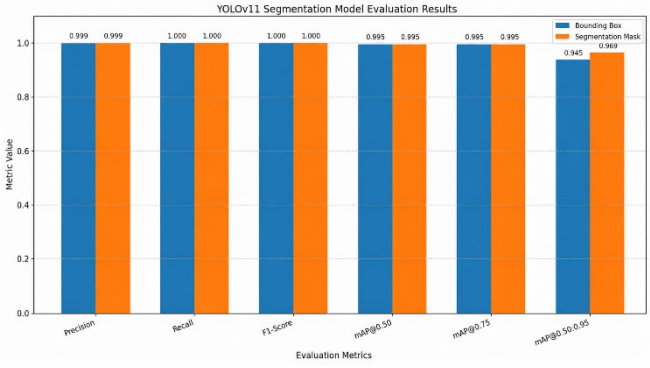


Figure 4. YOLOv11 segmentation results.

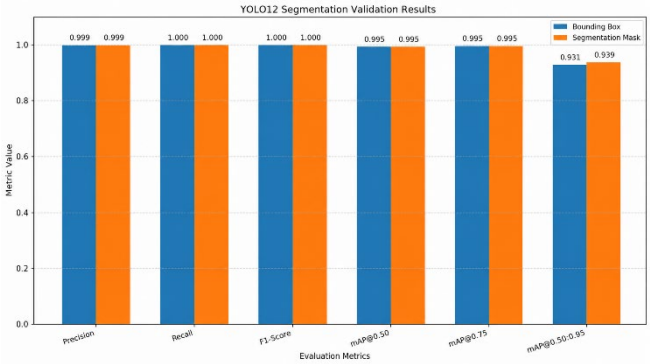


Figure 5. YOLOv12 segmentation results.

Figures 2–5 present the bounding-box and segmentation-mask metrics obtained by YOLOv8, YOLOv9, YOLOv11, and YOLOv12, respectively, on the independent test set. All models achieved similarly high precision, recall, F1 Score, and mAP@0.50. The principal differences between the models emerged for mAP@0.50:0.95, which incorporates more stringent overlap thresholds. Figure 2 shows that YOLOv8 achieved the best localization performance, with a bounding-box mAP@0.50:0.95 of 0.962750, while also providing the highest processing speed. Figure 3 demonstrates that YOLOv9 achieved the highest segmentation performance, with a mask mAP@0.50:0.95 of 0.970460. Figure 4 shows that YOLOv11 attained a mask mAP@0.50:0.95 of 0.969176, closely approaching YOLOv9, although its inference time was longer. Figure 5 indicates that YOLOv12 preserved high precision and perfect recall and operated at a speed comparable to YOLOv8; however, its bounding-box and mask mAP@0.50:0.95 values of 0.930811 and 0.938581, respectively, reflect lower localization performance under stringent overlap thresholds than the other evaluated models.

V. CONCLUSION

In this study, a comparative performance analysis of YOLOv8, YOLOv9, YOLOv11, and YOLOv12 was conducted for dental implant segmentation, and the findings demonstrated that each model offers distinct advantages in terms of accuracy and computational speed. The fact that all four models achieved an mAP@0.50 of 0.995 for both bounding boxes and segmentation masks and a recall of 1.000 indicates highly reliable implant detection on the test set. For the more discriminative mAP@0.50:0.95 metric, YOLOv8 achieved the highest bounding-box localization performance (0.962750), whereas YOLOv9 achieved the highest pixel-level segmentation performance for masks (0.970460). YOLOv11 attained a mask mAP@0.50:0.95 of 0.969176, closely approaching YOLOv9, but remained slower, with an inference time of 24.824682 ms and a total processing time of 34.000603 ms per image. YOLOv12 achieved a precision of 0.999062, a recall of 1.000, an F1-score of 0.999531, and approximately 92.324 FPS with an inference time of 10.831413 ms; however, its bounding-box and mask mAP@0.50:0.95 values were lower at 0.930811 and 0.938581, respectively.

YOLOv8's inference time of 10.309349 ms, total processing time of 15.968530 ms per image, and throughput of approximately 96.999 FPS indicate that it provides the most balanced option for clinical scenarios in which accuracy and real-time operation must be optimized jointly. Accordingly, YOLOv9 may be preferred in research and detailed analytical applications where maximum mask segmentation accuracy is prioritized, whereas YOLOv8 is more suitable when speed, bounding-box accuracy, and overall computational efficiency

must be considered together. From clinical and academic perspectives, the study contributes to strengthening decision-support systems, reducing variability associated with manual analysis, and providing a methodological framework that can be adapted to other medical imaging problems.

REFERENCES

[1]. Stuart C. White, M. J. (2013). Oral Radiology: Principles and Interpretation.
[2]. Elif Şener, B. O. (2023). The first step of artificial intelligence in dental practice: Segmentation applications. *Current Research in Dental Sciences*.
[3]. Salih Taha Alperen Özçelik, H. Ü. (2024). Derin öğrenme ile panoramik radyografi görüntülerinden diş segmentasyonu: UNet, FPN ve PSPNet karşılaştırması. *Niğde Ömer Halisdemir University Journal of Engineering Sciences*.
[4]. Yongye Su, Q. L. (2022). YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms—Computer *Methods and Programs in Biomedicine*.
[5]. Muhammed Enes ATIK, Z. D. (2021). Comparison of YOLO Versions for Object Detection from Aerial Images. *International Journal of Environment and Geoinformatics*.
[6]. Eslam Mohammed, A. S.-S. (2024). INSTA-YOLO: Real-Time Instance Segmentation. *arXiv preprint arXiv:2102.06777*.
[7]. Ali Badiezadeh, A. M. (2024). SEGMENTATION STRATEGIES IN DEEP LEARNING FOR PROSTATE CANCER DIAGNOSIS: A COMPARATIVE STUDY OF MAMBA, SAM, AND YOLO. *arXiv preprint arXiv:2409.16205*.
[8]. Eko Prasetyo, N. S. (2020). A Comparison of YOLO and Mask R-CNN for Segmenting the Head and Tail of Fish. *2020 4th International Conference on Informatics and Computational Sciences (ICICoS).* IEEE.
[9]. Cheng, R. (2019). A survey: Comparison between Convolutional Neural Network and YOLO in image identification. *Journal of Physics: Conference Series (Vol. 1453, No. 1, p. 012139).* IOP Publishing.
[10]. Yong-Hwan Lee, Y. K. (2020). Comparison of CNN and YOLO for Object Detection. *Journal of the Semiconductor & Display Technology*.
[11]. Halil Murat Ünver, E. A. (2019). Skin Lesion Segmentation in Dermoscopic Images with a Combination of YOLO and GrabCut Algorithm. *Diagnostics*.
[12]. Xiaoting Liang, X. J. (2022). Real-Time Grading of Defect Apples Using Semantic Segmentation Combination with a Pruned YOLO V4 Network. *Foods*.
[13]. Muhammad Shoaib, N. S. (2022). YOLO Object Detector and Inception-V3 Convolutional Neural Network for Improved Brain Tumor Segmentation. *Traitement Du Signal*.
[14]. Chang Ho Kang, S. Y. (2023). Real-time object detection and segmentation technology: an analysis of the YOLO algorithm. *JMST Advances*.
[15]. Bahaaeldeen M. Elgarba, S. V. (2023). Deep learning-based segmentation of dental implants on cone-beam computed tomography images: A validation study—Journal *of Dentistry*.
[16]. Tae Jun Jang, K. C. (2022). A Fully Automated Method for 3D Individual Tooth Identification and Segmentation in Dental CBCT. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*.
[17]. S. LEE, S. W. (2020). Automated CNN-Based Tooth Segmentation in Cone-Beam CT for Dental Implant Planning. *IEEEAccess*.
[18]. Monica Macrì, V. D.-F. (2024). The Role and Applications of Artificial Intelligence in Dental Implant Planning: A Systematic Review. *bioengineering*.

[19]. Yang Xing, P. L.-S. (2024). Dental X-ray Segmentation and Auto Implant Design Based on Convolutional Neural Network. *IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*.
[20]. Emel Gardiyanoğlu, G. Ü. (2023). Automatic Segmentation of Teeth, Crown–Bridge Restorations, Dental Implants, Restorative Fillings, Dental Caries, Residual Roots, and Root Canal Fillings on Orthopantomographs: Convenience and Pitfalls. *diagnostics*.
[21]. Mohammad Adel Moufti, N. T. (2023). Developing an Artificial Intelligence Solution to Autosegment the Edentulous Mandibular Bone for Implant Planning. *European Journal of Dentistry*.
[22]. Pedro Cunha, M. A. (2012). A method for segmentation of dental implants and crestal bone. *International Journal of Computer Assisted Radiology and Surgery*.
[23]. Ruben Pauwels, R. J. (2014). Automated implant segmentation in cone-beam CT using edge detection and particle counting—International *Journal of Computer Assisted Radiology and Surgery*.
[24]. Ultralytics, "YOLOv8: A state-of-the-art object detection model," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.
[25]. Jacob Solawetz, F., "What's new in YOLOv8," Roboflow, Nov. 27, 2024.
[26]. Chien-Yao Wang, I.-H. Y.-Y. (2024). YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *European Conference on Computer Vision*.
[27]. P. Potrimba, "What is new in YOLOv9? An architecture deep dive," Roboflow, Nov. 27, 2024.
[28]. Rahima Khanam, M. H. (2024). YOLOv11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.
[29]. S. N. Rao, "YOLOv11 architecture explained: Next-level object detection with enhanced speed and accuracy," Medium, Nov. 27, 2024.
[30]. Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," Advances in Neural Information Processing Systems, vol. 38, pp. 78433–78457, 2026.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

PhysioReco: An Evidence-Guided Machine Learning System for Personalized Arthritis Exercise Recommendation

Theresa Omodunbi
Dept of Information Systems
Obafemi Awolowo University,
Ile-Ife, Nigeria
tessydunbi@oauife.edu.ng
ORCID: 0000-0002-8501-6743

Temitope Sholanke
Dept of Computer Science and Cybersecurity
Obafemi Awolowo University,
Ile-Ife, Nigeria
tsholanke@oauife.edu.ng

Adesola Kareem
Dept of Computer Science and Cybersecurity
Obafemi Awolowo University,
Ile-Ife, Nigeria
kareemadesola1999@gmail.com

Grace Alilu
Dept of Information Systems
Obafemi Awolowo University,
Ile-Ife, Nigeria
gealilu@oauife.edu.ng

Adekemi Amoo
Dept of Software Engineering
Obafemi Awolowo University,
Ile-Ife, Nigeria
aamoo@oauife.edu.ng

Adaobi Okonji
Dept of Medical Rehabilitation
Obafemi Awolowo University,
Ile-Ife, Nigeria
amokonji@oauife.edu.ng

Abstract
This study presents the development and implementation of a data-driven physiotherapy recommender system empowered by machine learning. The system serves as a clinical decision-support tool for patients with arthritis, providing personalized exercise recommendations based on structured clinical parameters. The dataset was sourced from MIMIC-IV (Medical Information Mart for Intensive Care, Version 2.2), a large de-identified electronic health records database. It comprised 25,225 arthritis-related clinical encounter records representing 15,160 distinct patients. Diagnosis-based severity classification was applied to derive patient profiles, and exercise recommendation labels were assigned using evidence-based guidelines adapted from the American College of Rheumatology (ACR) and the Centers for Disease Control and Prevention (CDC). The resulting labeled dataset was partitioned into a training set of 20,180 records (80%) and a held-out test set of 5,045 records (20%) using stratified random sampling. A Random Forest Classifier with 200 decision trees was trained using six clinical features: age, sex, arthritis type, pain severity, mobility limitation, and use of a walking aid. The model achieved 100% accuracy and a macro AUC-ROC of 1.0000 across four exercise recommendation classes on the held-out test set. In addition, 20 representative system-generated recommendations were independently reviewed by a rehabilitation expert, and all were judged clinically appropriate. The system was deployed as a responsive Flask web application that provided personalized recommendations incorporating exercise type, frequency, intensity, and safety guidance. The high classification performance primarily reflects the deterministic nature of the rule-based label assignment and demonstrates the model's ability to reproduce the predefined recommendation logic rather than clinical effectiveness. Future work should validate the recommendations using prospectively collected, independently clinician-assigned labels and patient outcomes.

Keywords: Physiotherapy, Recommender System, Arthritis, Machine Learning, Electronic Health Records, MIMIC-IV, Healthcare, Personalized Exercise Recommendation



I. INTRODUCTION

Physiotherapy is a crucial component of rehabilitative healthcare, employing tailored exercise programs and specialized tools to address abnormal physical function. It addresses various health issues, including back pain, arthritis, osteoporosis, and nerve-related ailments [1]. Globally, approximately 2.4 billion people require rehabilitation [2], yet there is a substantial unmet need for these services, particularly in regions such as Nigeria, where the density of physiotherapists is alarmingly low [3], [4]. In Nigeria, the need for qualified physiotherapists, combined with economic constraints, limits access to quality physiotherapy services. The ratio of physiotherapists per population is critically low compared to developed countries, exacerbating the problem [3]. As a result, many individuals with potentially manageable conditions do not receive timely and effective treatment, leading to worsening health outcomes and increased healthcare costs. Arthritis is a disease that causes damage to your joints [5]. It is a common and debilitating condition that affects millions of people worldwide and significantly impairs quality of life. Effective management of arthritis often requires personalized exercise regimens tailored to alleviate specific symptoms [6], [7], [8], [9]. The current approach to physiotherapy in Nigeria typically relies on manual assessments and generalized treatment plans, often lacking the customization needed to address individual patient needs effectively. This traditional method is hampered by a lack of data-driven solutions, making it challenging to provide personalized care. Without integrating advanced technologies and data analytics, physiotherapy services remain largely one-size-fits-all, which may be ineffective for all patients, particularly those with complex or chronic conditions such as arthritis [10]. The development of a physiotherapy recommender system can bridge this gap by offering personalized exercise recommendations based on structured clinical parameters derived from patient health records [11], [12], [13], [14]. By leveraging machine learning and data analysis, such a system can enhance the precision and effectiveness of physiotherapy, ensuring that patients receive the most appropriate exercises for their specific conditions [15], [16], [17], [18]. The goal of this project is to create and implement a data-driven recommendation system for physical therapy, especially suited to people with arthritis. The technology uses machine learning algorithms trained on large-scale electronic health records to evaluate patient clinical profiles and provide tailored exercise recommendations. Enhancing patient outcomes, expanding access to efficient physiotherapy, and making the best use of the limited resources available to the healthcare industry are the goals of this research. To ensure the system meets the needs of both patients and healthcare providers, the study focuses on its design, development, and evaluation. This strategy not only addresses the immediate needs of arthritis patients but also advances broader initiatives to improve healthcare delivery in resource-constrained settings.

II. RELATED WORKS

By providing personalized exercise plans, Physiotherapy Recommender Systems (PRSs) have become increasingly significant in improving the care of patients with arthritis. These systems incorporate technologies such as fuzzy logic [17], [19], machine learning [15], [18], [20], [21], [22], and sensor data analysis [23], [24] to make such recommendations. Several important strategies and limitations are highlighted by a review of previous research, which is what this study aims to address. Fuzzy logic was employed in the early PRS implementations to suggest physical rehabilitation activities for stroke victims [17]. When dealing with uncertainty in patient-reported data—where symptoms aren't always well-defined—fuzzy logic was especially well-suited. These systems are difficult to create and improve, and the accuracy of their data is highly dependent on its quality. Inaccurate or incomplete patient profiles frequently led to inadequate exercise recommendations and unsatisfactory outcomes. Exercise recommendations are becoming increasingly personalized thanks to the growing use of machine learning techniques. For example, one study developed exercise programs using a multi-output regression model based on clinical information and patient-reported symptoms [18]. Although this approach made workouts more personalized, it faced challenges with scalability and real-time feedback. Subjectivity further affected the

consistency of the suggestions in patient-reported data. Another study used clinical evaluations and physical activity logs to create personalized exercise regimens for people with osteoarthritis using machine learning algorithms [15]. The method was constrained by a small sample size and a lack of long-term evaluation, despite showing improvements in pain management and physical function. To enhance real-time feedback and adherence tracking, some PRSs have combined machine learning with wearable sensor data [23]. These technologies provide real-time feedback on patient activities and tailored workout recommendations by evaluating sensor data. This strategy increased adherence rates and reduced pain. Still, it was difficult to fully evaluate its long-term effectiveness due to the studies' short intervention periods and the absence of control groups. Drug Recommender Systems (DRSs), which incorporate feedback mechanisms and characteristics relevant to individual patients, have also added to the body of knowledge. To deliver prescription recommendations based on patient profiles, DRSs frequently use machine-learning techniques like sentiment analysis and clustering [16]. Nevertheless, many DRSs fall short of incorporating real-time user feedback or considering age and prior medical conditions. As a result, more patient-centric PRSs with real-time symptom monitoring and feedback systems have been created. This work aims to overcome these constraints by creating a PRS that uses a Random Forest model to provide arthritis patients with tailored exercise recommendations. Compared to previous models, Random Forest has numerous advantages: it combines multiple decision trees to improve accuracy; it is robust by minimizing overfitting; and it handles missing data well. The flexibility of Random Forest enables reliable predictions and rankings of feature importance, which help identify the most influential clinical variables for exercise recommendations. This study directly addresses the scalability and data quality challenges that have hampered earlier systems by training and evaluating the model on a large-scale, real-world EHR dataset of 25,225 clinical records sourced from the MIMIC-IV database, substantially exceeding the sample sizes reported in prior work.

III. MATERIALS AND METHODS

This section describes the development process for the Physiotherapy Recommender System for patients with arthritis. It discusses the machine learning model's development, the datasets used, the system design, and the implementation environment. The strategy used ensures the system's dependability and functionality while providing precise exercise recommendations tailored to each patient's needs. Every part of the system, including data collection, preprocessing, model training, and evaluation, is covered in detail.

1. System Design

The study follows a well-defined architecture, as depicted in Figure 1, to provide exercise recommendations for patients with arthritis efficiently. The central components of this architecture include the User Interface (UI), Web Application, Machine Learning Model, and Data Management.

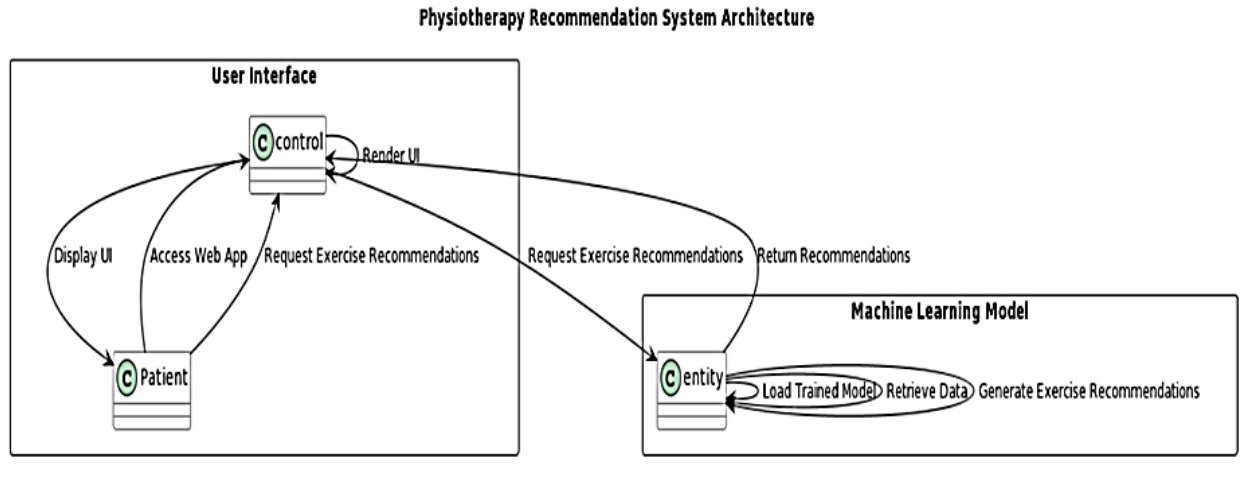


Figure 1. General Architecture for the Physiotherapy Recommender System

The User Interface serves as the front-end platform through which users interact with the system, presenting a structured clinical parameter input form and displaying personalized exercise recommendations. The Web Application serves as the intermediary, receiving user input and passing it to the Machine Learning Model. The core of the system is the Machine Learning Model, which uses a trained Random Forest Classifier to process clinical parameters and generate personalized exercise recommendations. The Data Management component ensures that model artifacts are efficiently persisted and loaded at application startup. This architecture harmonizes user interaction, data processing, and machine learning capabilities to deliver tailored exercise recommendations for individuals with arthritis, streamlining the process from clinical input to meaningful output.

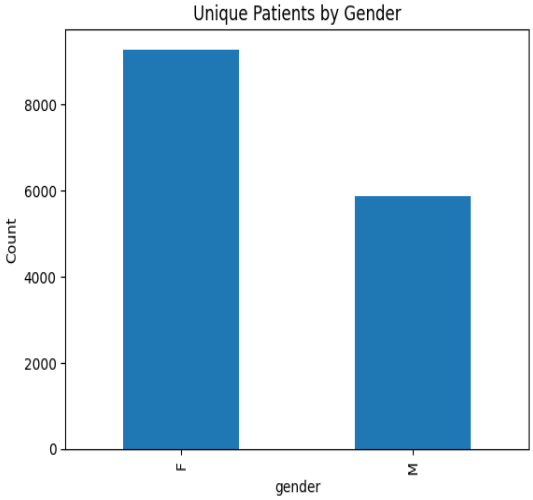
2. Dataset

The dataset used in this study was sourced from MIMIC-IV (Medical Information Mart for Intensive Care, Version 2.2), a large, de-identified electronic health records (EHR) database maintained by the Beth Israel Deaconess Medical Center (BIDMC) and publicly available through PhysioNet [25]. MIMIC-IV contains records for patients admitted to the BIDMC emergency department and intensive care units. Version 2.2 of the dataset used for this study comprises 431,231 records from patient admissions. These records are for 299,712 distinct patients with various ailments such as arthritis, diabetes, hypertension, peptic ulcer disease, mental disorders, neoplasms, etc. From this study, 25,225 admission records corresponding to 15,160 distinct patients with arthritis-related diagnoses were extracted. These records span 20 distinct arthritis variants, with the most prevalent being unspecified osteoarthritis (n = 7,879), rheumatoid arthritis (n = 7,029), and primary osteoarthritis of the knee and hip (n = 4,549). Exercise recommendation labels were not natively available in MIMIC-IV; they were derived using evidence-based labeling rules adapted from the Osteoarthritis Initiative (OAI) guidelines published by the ACR and CDC. Data Collection: Records were extracted from MIMIC-IV for all patients carrying a primary or secondary arthritis diagnosis. Each record provided patient age, sex, and ICD-10-coded diagnosis strings. The dataset comprised 25,225 arthritis-related clinical encounter records obtained from 15,160 distinct patients. Because the data were collected during routine hospital visits, some patients contributed multiple records for different clinical encounters. Each record was treated as an individual clinical encounter for the recommendation modeling task. The analysis of the dataset for patients diagnosed with arthritis is as follows:

i. Gender Distribution: A total of 9,279 females and 5,881 males were admitted for arthritis. The total number of admissions was 16272, with 8953 for females and 7319 for males. The numbers of males and females diagnosed with arthritis are shown in Figure 2.

ii. Age Distribution: Figure 3 shows that the average patient age is approximately 66 years, consistent with literature indicating that arthritis is elderly-dominated [5]. The standard deviation of the age distribution is approximately 15, indicating that patients are moderately spread across age groups. Not every patient is elderly, but most are clustered around older adults. The youngest patient in the dataset is 18 years old, and the oldest is 91 years old. This further indicates that the dataset includes only adults (no pediatric patients). The figure also shows that most patients are over 50 years of age, with the largest number in the 66-80 age range. This strongly aligns with the clinical prevalence of arthritis and related musculoskeletal conditions in older adults.

iii. Length of Hospital Stay (Admission Length): Table 1 shows information regarding the length of stay (LOS) per hospital admission. The LOS calculated is for 25,225 admissions. Therefore, every hospital record has a stay duration. The average length of hospital stay is approximately 5 days. This suggests that mostly short-to-moderate hospitalizations.



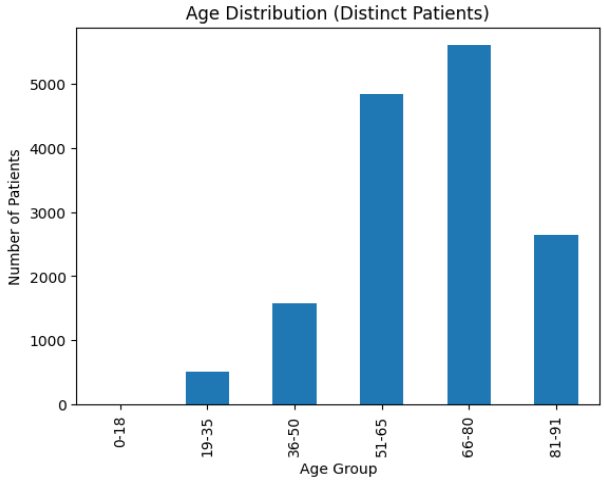


Figure 2. Unique diagnoses by gender, Figure 3. Age Distribution (For Distinct Patients)

Table 1. Length of Hospital Stay

	Length of Hospital Stay
count	25,225
mean	5.35
std	7.69
min	1
25%	1
50%	3
75%	6
max	200

The LOS standard deviation of approximately 8 is quite high relative to the mean. This means that the LOS is highly variable, as some patients stay very short while others stay very long.

25% of patients stay 1 day or less. This can be due to many short admissions, including emergency or observation cases. Half of the patients stay 3 days or less. 75% of patients stay 6 days or less. So most admissions are under a week.

The maximum number of days of stay is 200. This means that some patients stayed approximately 6–7 months. This can be as a result of severe chronic illness, complications, long-term hospitalization in ICU or recovery, etc.

iv. Number of Deaths: 3,450 of the patients diagnosed with arthritis died according to the dataset. This accounts for 22.76% of all arthritis patients. Some of these patients died during admission, others after discharge, and some during home healthcare or at a skilled nursing facility. Figure 4 is a bar chart showing patient survival vs. death. The first bar has the label Alive/Unknown because some of the patients might have died after discharge without the hospital being aware.

To assign exercise labels, each diagnosis was classified into one of five severity tiers—at risk, mild, moderate, moderate-severe, and severe—using keyword-based rules applied to the diagnosis text. Exercise labels, frequency prescriptions, intensity levels, and safety notes were then assigned according to four evidence-based labeling rules (R1–R4) adapted from OAI-style guidelines, as summarized in Table 2. Each rule maps a severity tier to an exercise recommendation, frequency, and intensity prescription.

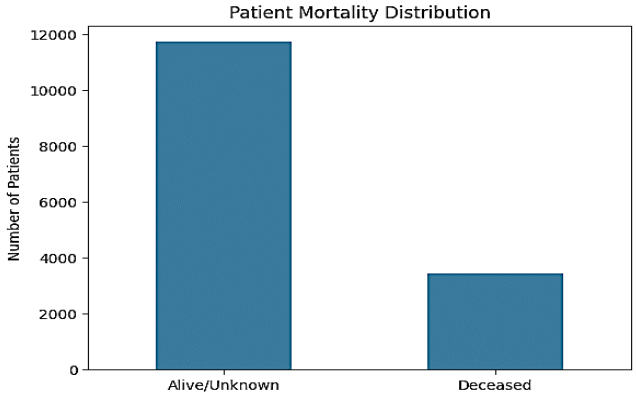


Figure 4. Survival Vs Death Plot

Table 2. Evidence-based exercise labeling rules

Rule	Severity Tier	Input Conditions	Recommended Exercise	Frequency	Intensity
R1	At risk / Mild	Pain: Mild; Mobility: Low; No walking aid	Walking/brisk walking + stretching	20–30 min, 3–5 days/week	Low to moderate

R2	Moderate	Pain: Moderate; Mobility: Moderate	Cycling/walking + light strengthening	20 min, 3 days/week	Low
R3	Moderate-Severe	Pain: Moderate-Severe; Mobility: High	Water exercise + tai chi/balance + gentle strengthening	15–25 min, 2–3 days/week	Low
R4	Severe	Pain: Severe or walking aid = Yes	Supervised aquatic/chair-based exercise + physiotherapy review	10–20 min, 2–3 days/week	Very low to low

Data Preprocessing: The following preprocessing steps were applied to the extracted records:

Severity Classification: Each ICD-10 diagnosis string was mapped to one of five severity tiers using priority keyword rules. Diagnoses containing 'rheumatoid' or 'pyogenic arthritis' map to the severe tier. Diagnoses containing 'bilateral' or 'polyosteoarthritis' map to moderate-severe. Diagnoses containing 'unilateral primary osteoarthritis' or 'primary osteoarthritis' of a specific joint map to moderate. Diagnoses containing 'hyperuricemia without signs of inflammatory arthritis' map to at-risk. All remaining osteoarthritis diagnoses default to mild. Where a record carries multiple diagnoses, the highest-priority tier takes precedence.

Feature Derivation: Six model input features were derived from the MIMIC-IV records: patient age (continuous, years); sex (binary: F/M); arthritis type (categorical: Knee OA, Hip OA, Rheumatoid Arthritis, Other OA, Other), derived from diagnosis keywords; pain severity (ordinal: Mild, Moderate, Moderate-Severe, Severe), derived from severity tier; mobility limitation (ordinal: Low, Moderate, High), derived from severity tier; and use of a walking aid (binary: Yes/No), inferred as 'Yes' for severe-tier patients aged 70 years or above and for moderate-severe-tier patients aged 75 or above, and 'No' otherwise.

Label Assignment: Exercise recommendation labels were assigned using four evidence-based rules (R1–R4). R1 assigns 'Walking/brisk walking + stretching' (20–30 min, 3–5 days/week, low to moderate intensity) to at-risk and mild-tier patients. R2 assigns 'Cycling/walking + light strengthening' (20 min, 3 days/week, low intensity) to moderate-tier patients. R3 assigns 'Water exercise + tai chi/balance + gentle strengthening' (15–25 min, 2–3 days/week, low intensity) to patients in the moderate-to-severe tier. R4 assigns 'Supervised aquatic/chair-based exercise + physiotherapy review' (10–20 min, 2–3 days/week, very low to low intensity) to severe-tier patients. Each rule also specifies a tailored safety note.

The exercise recommendation labels were not directly available in the MIMIC-IV database and were therefore generated using predefined evidence-based rules (R1–R4). These rules derive the recommendation class from clinical characteristics such as pain severity, mobility limitation, and walking-aid status. Several of these characteristics are also included among the model input features. Therefore, the resulting machine-learning task evaluates the Random Forest's ability to reproduce the predefined rule-based recommendation logic rather than predict independently observed clinical outcomes.

Feature Encoding: Categorical features were encoded using a deterministic ordinal scheme: pain severity (Mild = 0, Moderate = 1, Moderate-Severe = 2, Severe = 3), mobility limitation (Low = 0, Moderate = 1, High = 2), sex (F = 0, M = 1), and walking aid use (No = 0, Yes = 1). Age was retained as a continuous numeric feature. The exercise label was encoded as an integer class index (0–3) corresponding to rules R1–R4, respectively.

Data Partitioning: The full dataset of 25,225 clinical encounter records (admissions records) was partitioned into a training set of 20,180 records (80%) and a held-out test set of 5,045 records (20%) using stratified random sampling to preserve the distribution of the four exercise recommendation classes. The partitioning was performed at the clinical encounter-record level, consistent with the modeling unit used in this study. Because some patients contributed multiple hospital encounter records, records belonging to the same patient could potentially occur in both the training and test subsets. Therefore, the held-out test set represents unseen clinical encounter records rather than exclusively unseen patients. This distinction is important when interpreting model performance, particularly with respect to generalization to new patients.

The four exercise classes in the training set are distributed as follows: Walking/brisk walking + stretching (n = 7,680), Supervised aquatic/chair-based exercise + physiotherapy review (n = 7,398), Cycling/walking + light strengthening (n = 3,914), and Water exercise + tai chi/balance + gentle strengthening (n = 1,188). Class-weight balancing was applied during model training to account for this imbalance.

This stratified partitioning ensures that the model is evaluated on genuinely unseen patient records and that minority exercise classes are adequately represented in both training and evaluation.

3. Exercise Recommender Model Formulation

The Random Forest Classifier [26], which forms the machine learning component, is encapsulated in the PhysioPredictor class. This model is well-suited to multi-class classification problems such as mapping a patient's clinical profile to an appropriate exercise recommendation. During training, the algorithm builds an ensemble of 200 decision trees, each trained on a random bootstrap sample of the data with a random subset of features considered at each node split. The majority vote across all trees determines the final class prediction, reducing variance and improving generalization relative to a single decision tree.

Mathematical Method:

Given a set of training examples $(x1,y1),(x2,y2),\ldots,(xn,yn)$ where xi represents the feature vector (symptoms) and yi represents the output, which is the exercise recommendation:

Bootstrap Sampling: Randomly select a subset of the training data with replacement.

Decision Tree Construction: For each bootstrap sample, a decision tree hk(x) is grown using a random subset of features at each node.

Voting: Each tree hk(x) in the forest gives a prediction, and the final output f(x) is determined by majority voting:

$$f(x) = \text{mode}\{h1(x), h2(x)\ldots, hB(x)\}$$

where B is the total number of trees in the forest.

The PhysioPredictor class encapsulates the machine learning component. Upon instantiation, it loads the serialized model artifact from disk. The class exposes a single predict() method that accepts six clinical parameters—age, sex, arthritis type, pain severity, mobility limitation, and use of a walking aid—encodes them deterministically, and returns the predicted exercise label together with its associated frequency, intensity, and safety note.

IV. RESULTS AND DISCUSSION

The updated Physiotherapy Recommender System features a structured clinical input form through which users enter six parameters: age, sex, arthritis type, pain severity, mobility limitation, and use of a walking aid. Based on these inputs, the system generates a personalized exercise recommendation in real time, along with the recommended frequency and intensity level, and a safety note tailored to the patient's severity profile. The updated interface replaces the original symptom-checkbox approach with validated clinical dropdowns, improving clinical relevance and enabling use of the substantially larger MIMIC-IV dataset. The system is deployed as a live web application accessible at: https://kareemadesola-physio-recommender.hf.space. By presenting inputs in structured clinical terms and providing multi-faceted recommendations, the system supports both patients and caregivers in selecting appropriate exercise regimens. The real-time recommendation engine, combined with severity-stratified safety guidance, ensures that users receive clinically aligned advice appropriate to their condition severity. The user-friendly design ensures that patients, regardless of technical literacy, can interact effectively with the system. Inputs such as walking aid use and mobility limitation help stratify patients accurately, and the system's output explicitly includes frequency and intensity alongside safety warnings, promoting safe adherence to the prescribed exercise plan.

1. System Interface

The system presents users with a structured clinical input form comprising six patient parameters: age (in years), biological sex (Female/Male), arthritis type (Knee OA, Hip OA, Rheumatoid Arthritis, Other OA, or Other), pain severity (Mild, Moderate, Moderate-Severe, or Severe), mobility limitation (Low, Moderate, or High), and use of a walking aid (Yes/No). Each parameter maps directly to an input feature of the trained Random Forest Classifier. Dropdown menus and radio buttons are used for categorical inputs to ensure data entry consistency and minimize user error. Upon form submission, the system immediately returns a personalized exercise recommendation comprising the exercise type, recommended frequency, intensity level, and a severity-stratified safety note. The interface is built with Bootstrap 5 and is fully responsive on desktop and mobile devices. Figure 5 shows the clinical input form, and Figure 6 shows a sample recommendation output for a 72-year-old female patient with Rheumatoid Arthritis (severe pain, high mobility limitation, uses a walking aid), for whom the system recommends supervised aquatic/chair-based exercise at 10–20 minutes, 2–3 days per week, at very low to low intensity.

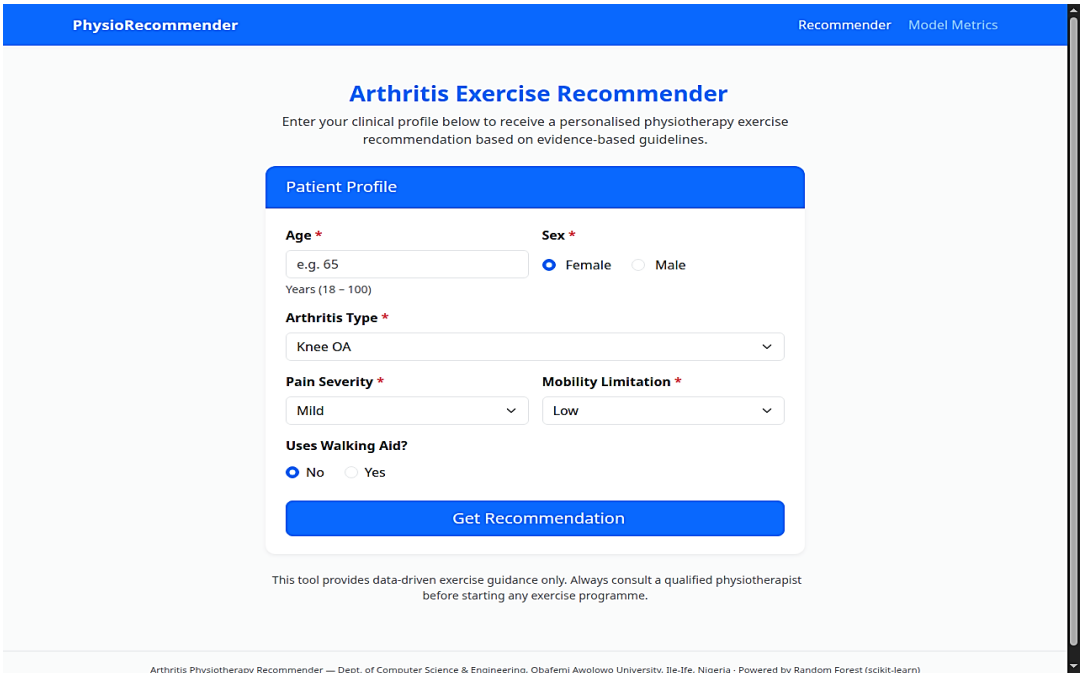


Figure 5. The PhysioReco clinical input form

2. System Output

The system generates exercise recommendations based on the six clinical inputs provided, as seen in Figures 5 and 6. For example, a 72-year-old female with Rheumatoid Arthritis, severe pain, high mobility limitation, and use of a walking aid receives a recommendation for 'Supervised aquatic/chair-based exercise + physiotherapy review' at a frequency of 10–20 minutes, 2–3 days per week, at very low to low intensity, with a safety note advising clinical review before initiating an independent program. A 55-year-old male with Knee Osteoarthritis, mild pain, and low mobility limitation receives 'Walking/brisk walking + stretching' at 20–30 minutes, 3–5 days per week, at low to moderate intensity. A 63-year-old female with Hip Osteoarthritis, moderate pain, and moderate mobility limitation receives 'Cycling/walking + light strengthening' at 20 minutes, 3 days per week. A 70-year-old female with bilateral Knee OA, moderate-to-severe pain, and significant mobility limitation receives 'Water exercise + tai chi/balance + gentle strengthening' for 15–25 minutes, 2–3 days per week. These examples demonstrate the system's capacity to distinguish among four clinically meaningful exercise classes along the OA severity spectrum. The results confirm the system's capacity to map structured clinical inputs to appropriate, severity-stratified exercise recommendations. The four exercise classes reflect a clinically ordered spectrum from low-impact aerobic activity for mild cases to fully supervised aquatic and chair-based therapy for severe cases. Transitioning from symptom-checkbox input to structured clinical parameters substantially improves the specificity, clinical interpretability, and scalability of the recommendation pipeline.

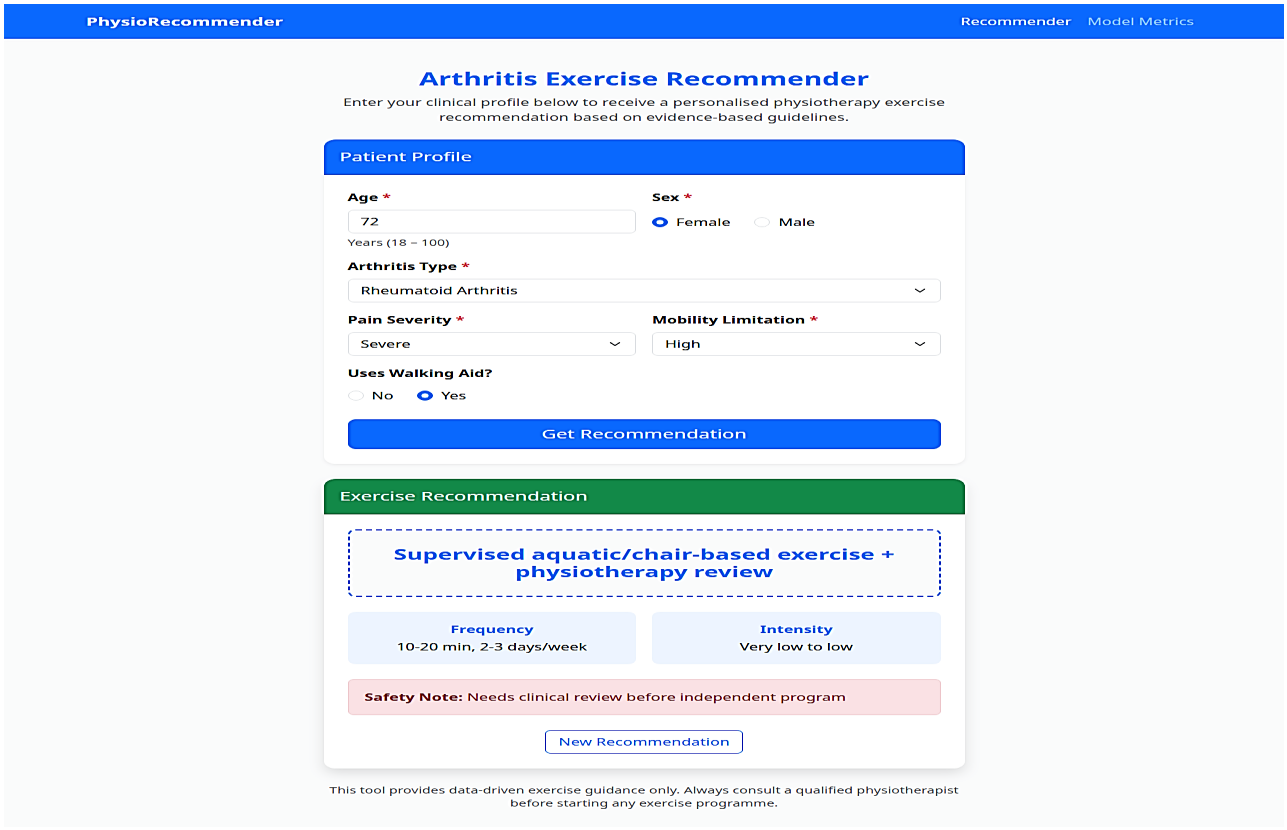


Figure 6. PhysioReco Output

3. System Testing and Validation

The system was evaluated on a held-out test set of 5,045 records (20% of the total 25,225-record dataset), selected using stratified random sampling to maintain the class distribution of the training data. A confusion matrix was computed across all four exercise recommendation classes [27], and AUC-ROC was computed using a one-vs-rest macro-averaged approach, as recommended for multi-class classification [28]. The normalized confusion matrix is presented in Figure 7. The rows represent the actual classes, while the columns represent the classes predicted by the Random Forest model. All diagonal entries have a value of 1.00, whereas all off-diagonal entries have a value of 0.00. This indicates that every test record was assigned to its correct exercise class, with no misclassifications across the four categories. The per-class classification report is summarised in Table 3. The per-class test support was: Walking/brisk walking + stretching (n = 1,920), Cycling/walking + light strengthening (n = 978), Water exercise + tai chi/balance + gentle strengthening (n = 297), and Supervised aquatic/chair-based exercise + physiotherapy review (n = 1,850). The model achieved 100% accuracy, with a precision, recall, and F1-score of 1.0000 across all four recommendation classes. However, this result must be interpreted in the context of the label-generation procedure. The exercise recommendation labels were deterministically generated using predefined evidence-based rules, and several variables used in these rules were also included as model input features. Therefore, the classification task primarily assesses the Random Forest's ability to reproduce the predefined recommendation logic. The 100% accuracy demonstrates complete agreement between the model predictions and the rule-derived labels in the held-out evaluation records. Still, it does not establish that the model can predict independently observed clinical outcomes with perfect accuracy or that the recommendations are clinically effective.

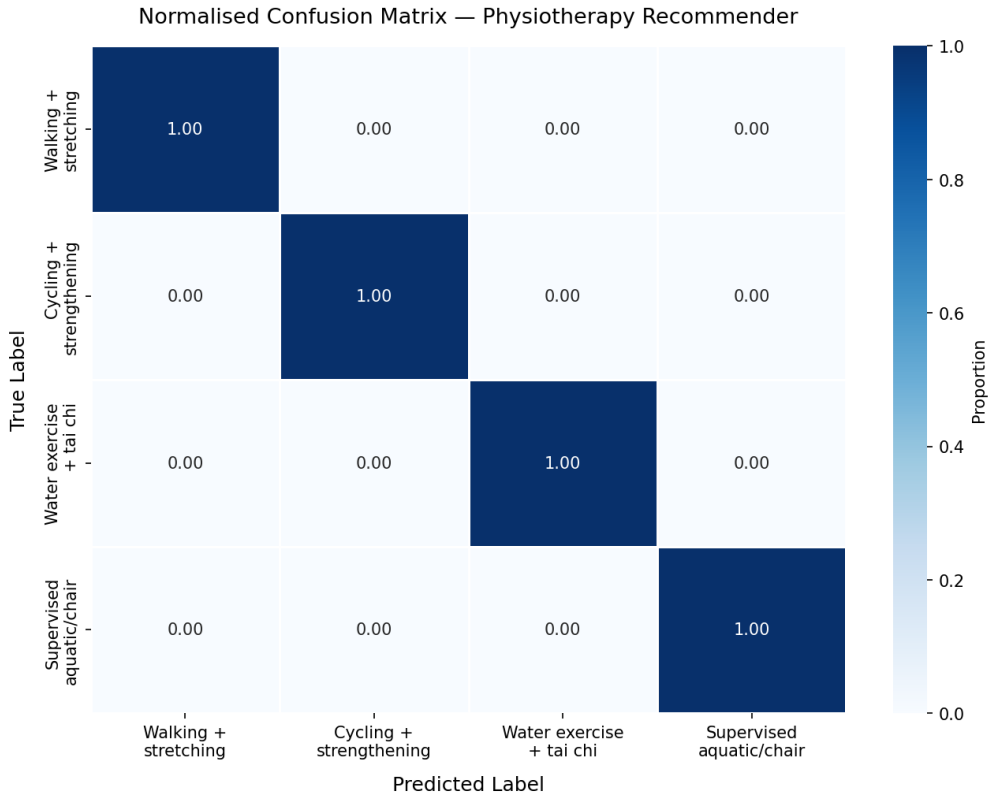


Figure 7. Normalized Confusion Matrix

Table 3. Exercise classification report evaluated

Exercise Class	Precision	Recall	F1-Score	Support (n)
Walking/brisk walking + stretching	1.0000	1.0000	1.0000	1,920
Cycling/walking + light strengthening	1.0000	1.0000	1.0000	978
Water exercise + tai chi/balance + gentle strengthening	1.0000	1.0000	1.0000	297
Supervised aquatic/chair-based exercise + physiotherapy review	1.0000	1.0000	1.0000	1,850
Macro Average	1.0000	1.0000	1.0000	5,045

4. Clinical Validation of Exercise Recommendations

To provide an independent clinical assessment of the recommendations generated by the PhysioReco system, an author with expertise in medical rehabilitation conducted a clinical validation exercise. The purpose of the validation was to determine whether the exercise recommendations generated by the system were clinically appropriate and consistent with established rehabilitation principles for patients presenting with different arthritis characteristics and levels of functional limitation. This assessment was conducted separately from the machine-learning performance evaluation because the 100% classification accuracy reported in the preceding section reflects agreement with rule-derived labels and does not, by itself, establish clinical appropriateness or effective-

ness. Twenty representative patient profiles were generated from the system using combinations of the six clinical input variables: age, sex, arthritis type, pain severity, mobility limitation, and use of a walking aid. The profiles were selected to cover the range of arthritis types and severity categories represented by the four exercise recommendation classes. Each profile was entered into the deployed PhysioReco system, and the rehabilitation expert independently reviewed the resulting recommendation, frequency, intensity, and safety guidance. The clinical validation results are presented in Table 4. Each recommendation was assessed as either clinically appropriate or inappropriate, with the option to provide comments when modification was deemed necessary. The rehabilitation expert judged all 20 reviewed recommendations to be clinically appropriate, corresponding to an expert agreement rate of 100% (20/20). The expert assessment, therefore, provided preliminary evidence that the system-generated recommendations were clinically plausible and appropriately aligned with the severity and functional characteristics of the reviewed cases. This assessment is complementary to the computational evaluation. Whereas the latter assesses reproduction of the rule-derived labels, the clinical review assesses the appropriateness of the resulting recommendations from a rehabilitation perspective.

Table 4. Clinical Validation of PhysioReco Recommendations by a Rehabilitation Expert

Case	Age	Sex	Arthritis Type	Pain Severity	Mobility Limitation	Walking Aid	System-Generated Recommendation	Expert Assessment
1	55	Male	Knee OA	Mild	Low	No	Walking/brisk walking + stretching; 20–30 min, 3–5 days/week; low–moderate intensity	Appropriate
2	61	Female	Knee OA	Mild	Low	No	Walking/brisk walking + stretching; 20–30 min, 3–5 days/week; low–moderate intensity	Appropriate
3	58	Female	Hip OA	Mild	Low	No	Walking/brisk walking + stretching; 20–30 min, 3–5 days/week; low–moderate intensity	Appropriate
4	66	Male	Other OA	Mild	Low	No	Walking/brisk walking + stretching; 20–30 min, 3–5 days/week; low–moderate intensity	Appropriate
5	63	Female	Knee OA	Moderate	Moderate	No	Cycling/walking + light strengthening; 20 min, 3 days/week; low intensity	Appropriate
6	68	Male	Hip OA	Moderate	Moderate	No	Cycling/walking + light strengthening; 20 min, 3 days/week; low intensity	Appropriate
7	59	Female	Other OA	Moderate	Moderate	No	Cycling/walking + light strengthening; 20 min, 3 days/week; low intensity	Appropriate
8	72	Female	Knee OA	Moderate	Moderate	No	Cycling/walking + light strengthening; 20 min, 3 days/week; low intensity	Appropriate
9	70	Female	Knee OA	Moderate-Severe	High	No	Water exercise + tai chi/balance + gentle strengthening; 15–25 min, 2–3 days/week; low intensity	Appropriate

10	74	Male	Hip OA	Moder-ate-Se-vere	High	No	Water exercise + tai chi/balance + gentle strengthening; 15–25 min, 2–3 days/week; low intensity	Ap-pro-priate
11	67	Fe-male	Other OA	Moder-ate-Se-vere	High	No	Water exercise + tai chi/balance + gentle strengthening; 15–25 min, 2–3 days/week; low intensity	Ap-pro-priate
12	75	Fe-male	Knee OA	Moder-ate-Se-vere	High	Yes	Water exercise + tai chi/balance + gentle strengthening; 15–25 min, 2–3 days/week; low intensity	Ap-pro-priate
13	72	Fe-male	Rheu-matoid Arthritis	Severe	High	Yes	Supervised aquatic/chair-based exercise + physiotherapy review; 10–20 min, 2–3 days/week; very low–low intensity	Ap-pro-priate
14	78	Male	Rheu-matoid Arthritis	Severe	High	Yes	Supervised aquatic/chair-based exercise + physiotherapy review; 10–20 min, 2–3 days/week; very low–low intensity	Ap-pro-priate
15	70	Fe-male	Rheu-matoid Arthritis	Severe	High	Yes	Supervised aquatic/chair-based exercise + physiotherapy review; 10–20 min, 2–3 days/week; very low–low intensity	Ap-pro-priate
16	81	Fe-male	Knee OA	Severe	High	Yes	Supervised aquatic/chair-based exercise + physiotherapy review; 10–20 min, 2–3 days/week; very low–low intensity	Ap-pro-priate
17	76	Male	Hip OA	Severe	High	Yes	Supervised aquatic/chair-based exercise + physiotherapy review; 10–20 min, 2–3 days/week; very low–low intensity	Ap-pro-priate
18	64	Fe-male	Rheu-matoid Arthritis	Severe	High	Yes	Supervised aquatic/chair-based exercise + physiotherapy review; 10–20 min, 2–3 days/week; very low–low intensity	Ap-pro-priate
19	69	Male	Knee OA	Moder-ate-Se-vere	High	No	Water exercise + tai chi/balance + gentle strengthening; 15–25 min, 2–3 days/week; low intensity	Ap-pro-priate
20	73	Fe-male	Hip OA	Moder-ate	Moder-ate	No	Cycling/walking + light strength-ening; 20 min, 3 days/week; low intensity	Ap-pro-priate

5. Model Performance and Recommendations

The Random Forest Classifier was configured with 200 decision trees, balanced class weights (to account for the imbalanced distribution across exercise classes), and a fixed random seed (42) for full reproducibility. The model was trained on 20,180 records using six input features: age, sex, arthritis type, pain severity, mobility limitation, and use of a walking aid. Feature importance analysis (Figure 8) reveals that pain severity was the

most influential feature in reproducing the rule-based exercise recommendation classes, accounting for approximately 47% of the total feature importance, followed by mobility limitation at approximately 34%. Arthritis type contributed approximately 16%, while walking-aid use had a comparatively small contribution. Age and sex had minimal influence on the classification outcome. These findings indicate that the model primarily distinguished among the four exercise recommendation classes using the severity and functional-status variables. This pattern is consistent with the labeling framework, in which pain severity and mobility limitation were the principal factors used to allocate patients to progressively more supported exercise interventions. The dominance of pain severity and mobility limitation is clinically understandable because these variables directly reflect a patient's capacity to participate safely in physical activity. Patients with mild pain and low mobility restrictions were generally assigned walking, stretching, and low- to moderate-intensity exercises. In contrast, those with severe pain or high mobility restriction were assigned supervised aquatic or chair-based exercises. Their high feature-importance values therefore confirm that the Random Forest reproduced the predetermined labeling logic rather than proving that these variables are independently the strongest predictors of successful physiotherapy outcomes. The model's performance on the MIMIC-IV-derived dataset represents a substantial advance over the original 30-record prototype, which was too small for statistically meaningful evaluation. The current model was trained on 20,180 clinical encounter records and evaluated on 5,045 held-out encounter records, providing an assessment of classification fidelity at the clinical encounter level. The shift from symptom-severity weights to structured clinical features (arthritis type, pain severity, mobility limitation) also improves the clinical interpretability of the model's decision boundaries. It aligns the recommendation logic with internationally recognized arthritis management guidelines.

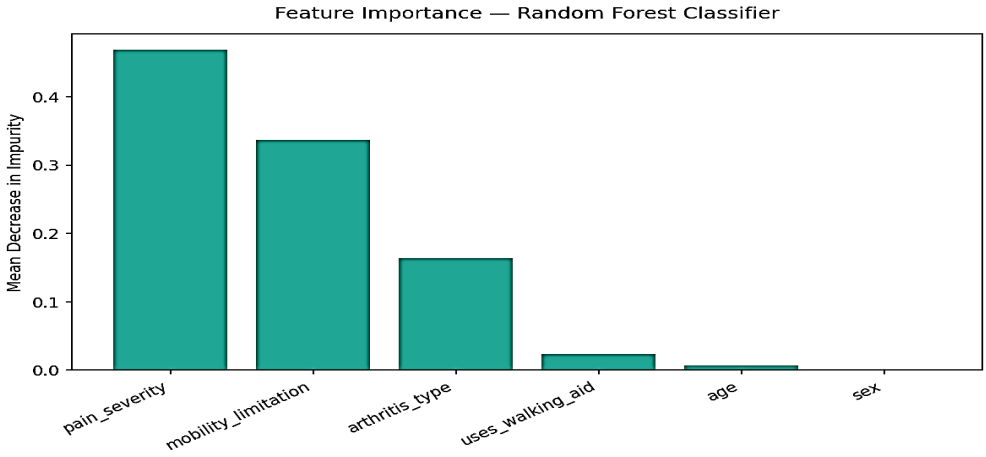


Figure 8. Feature Importance scores

V. CONCLUSION AND RECOMMENDATION

The Arthritis Physiotherapy Recommender System presents a scalable, data-driven tool for personalized exercise guidance for patients with arthritis. Built on 25,225 de-identified clinical records (distinct admission records) from the MIMIC-IV database and trained using evidence-based exercise labeling rules adapted from ACR and CDC guidelines, the system demonstrates that a Random Forest Classifier can reliably learn and replicate structured clinical decision logic for exercise prescription across four clinically ordered recommendation classes. The model achieved perfect accuracy and AUC-ROC on a 5,045-record held-out test set, and the system is deployed as a responsive Flask web application that provides real-time recommendations, including exercise type, frequency, intensity, and safety guidance. While the deterministic nature of the rule-based labeling scheme limits

the independence of the current evaluation, the architecture—which uses a large, real-world EHR dataset and validated clinical guidelines—provides a robust foundation for future work that incorporates prospectively labeled clinical data, richer feature sets, and real-world validation in clinical practice. The result confirms that the implementation correctly captures the intended decision rules, but further evaluation using prospectively collected, clinician-labeled data is required to establish external validity and clinical effectiveness. A limitation of the current study is that the dataset was partitioned at the admission record level rather than at the patient level. As a result, when patients contributed multiple hospital visits, records from the same patient could appear in both the training and evaluation subsets. Therefore, the reported performance should be interpreted as performance at the admission-record level rather than as authoritative evidence of generalization to completely unseen patients. Future studies should employ patient-level partitioning and external validation to provide a more stringent assessment of generalizability. Also, although the model achieved 100% accuracy in reproducing the evidence-guided recommendation labels, the labels were derived from predefined rules rather than independently observed physiotherapy outcomes. As a result, the predictive evaluation cannot establish clinical effectiveness. The expert-based evaluation provides preliminary evidence of clinical appropriateness but does not constitute prospective clinical validation or demonstrate improved patient outcomes. Future studies should evaluate the system prospectively with multiple physiotherapists and arthritis patients and compare recommendations with longitudinal treatment outcomes. Future improvements to the system should include: (1) collection of prospectively labeled datasets with physiotherapist-assigned exercise labels to enable independent clinical validation; (2) integration of additional variables such as Body Mass Index (BMI), WOMAC pain and function scores, and Kellgren-Lawrence (KL) grade to enrich the clinical feature space; (3) incorporation of patient feedback and exercise adherence tracking to support continuous model refinement; (4) deployment in a clinical pilot study, with physiotherapist oversight, to establish clinical validity, real-world utility, and patient outcomes in an African healthcare context, (5) prospectively labeled patient data should be collected where exercise labels are assigned by certified physiotherapists—to enable a rigorous, independent assessment of model generalization to real clinical populations. Since age is continuous and the other variables are encoded categorically, permutation importance or SHAP analysis could be explored in future work to provide a more reliable and clinically interpretable explanation of model predictions.

AI DISCLOSURE

During manuscript preparation, generative artificial intelligence tools were used only for language refinement, organization of ideas, grammar correction, and readability improvement. The tools were not used to generate or alter the study data, perform statistical analysis, train or evaluate the machine-learning models, create clinical interpretations, or make authorship-level contributions. All authors reviewed, verified, and approved the final manuscript and take full responsibility for the accuracy, integrity, originality, and scientific content of the work.

REFERENCES

[1]. S. O'Hanlon and M. Smith, A Comprehensive Guide to Rehabilitation of the Older Patient E-Book: A Comprehensive Guide to Rehabilitation of the Older Patient E-Book. Elsevier Health Sciences, 2020. Accessed: Aug. 30, 2024. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=lnEJEAAAQBAJ&oi=fnd&pg=PP1&dq=O%E2%80%99Reilly,+F.+(2020).+WHAT+IS+PHYSIOTHERAPY%3F.+A+Comprehensive+Guide+to+Rehabilitation+of+the+Older+Patient+E-Book:+A+Comprehensive+Guide+to+Rehabilitation+of+the+Older+Patient+E-Book,+118,+59.&ots=NUPRieW_TP&sig=2DGaIMu_1Wmlsi8JfYXKtEqAVJw

[2]. J. Bickenbach, C. Sabariego, and G. Stucki, "Beneficiaries of rehabilitation," Archives of physical medicine and rehabilitation, vol. 102, no. 3, pp. 543–548, 2021.
[3]. I. J. Odumodu, T. F. Olufunlayo, B. E. Ogunnowo, and M. E. Kalu, "Satisfaction With Services Among Attendees of Physiotherapy Outpatient Clinics in Tertiary Hospitals in Lagos State," Journal of Patient Experience, vol. 7, no. 4, pp. 468–478, Aug. 2020, doi: 10.1177/2374373519847370.
[4]. "World Bank Open Data," World Bank Open Data. Accessed: Aug. 30, 2024. [Online]. Available: https://data.worldbank.org
[5]. "Arthritis: Symptoms, Causes, Types, Treatment & Prevention." Accessed: Sep. 01, 2024. [Online]. Available: https://my.clevelandclinic.org/health/diseases/12061-arthritis
[6]. A. M. Davis, K. D. Davis, S. T. Skou, and E. M. Roos, "Why Is Exercise Effective in Reducing Pain in People with Osteoarthritis?" Curr Treat Options in Rheum, vol. 6, no. 3, pp. 146–159, Sep. 2020, doi: 10.1007/s40674-020-00154-x.
[7]. L. Anggiat, A. J. Altavas, and W. Budhyanti, "Joint Mobilization: Theory and evidence review," International Journal of Sport, Exercise and Health Research, vol. 4, no. 2, pp. 86–90, 2020.
[8]. S. Thiengwittayaporn et al., "Development of a mobile application to improve exercise accuracy and quality of life in knee osteoarthritis patients: a randomized controlled trial," Arch Orthop Trauma Surg, vol. 143, no. 2, pp. 729–738, Aug. 2021, doi: 10.1007/s00402-021-04149-8.
[9]. L. N. Medrado, M. L. M. Mendonça, M. B. Budib, S. A. Oliveira-Junior, and P. F. Martinez, "Effectiveness of aquatic exercise in the treatment of inflammatory arthritis: systematic review," Rheumatol Int, vol. 42, no. 10, pp. 1681–1691, May 2022, doi: 10.1007/s00296-022-05145-w.
[10]. D. Akeju et al., "The use of and needs for data and information by health professionals delivering palliative cancer care in sub-Saharan Africa: A qualitative study," Health Informatics Journal, vol. 28, no. 4, 2022, Accessed: Sep. 01, 2024. [Online]. Available: https://eprints.whiterose.ac.uk/192516/
[11]. T. N. T. Tran, A. Felfernig, C. Trattner, and A. Holzinger, "Recommender systems in the healthcare domain: state-of-the-art and research issues," J Intell Inf Syst, vol. 57, no. 1, pp. 171–201, Aug. 2021, doi: 10.1007/s10844-020-00633-6.
[12]. U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, and S. Luo, "A review of content-based and context-based recommendation systems," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 3, pp. 274–306, 2021.
[13]. R. De Croon, L. Van Houdt, N. N. Htun, G. Štiglic, V. V. Abeele, and K. Verbert, "Health recommender systems: systematic review," Journal of Medical Internet Research, vol. 23, no. 6, p. e18035, 2021.
[14]. P. Howell, A. Aryal, and C. Wu, "Web-Based Patient Recommender Systems for Preventive Care: Protocol for Empirical Research Propositions," JMIR Research Protocols, vol. 12, no. 1, p. e43316, 2023.
[15]. L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou, and D. Chen, "Research on machine learning algorithms and feature extraction for time series," in 2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC), IEEE, 2017, pp. 1–5. Accessed: Sep. 01, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8292668/
[16]. T. O. Omodunbi, G. E. Alilu, and R. N. Ikono, "Drug Recommender Systems: A Review of State-of-the-Art Algorithms," 2022 5th Information Technology for Education and Development (ITED), pp. 1–8, 2022.
[17]. C. Gmez-Portes, J. J. Castro-Schez, J. Albusac, D. N. Monekosso, and D. Vallejo, "A fuzzy recommendation system for the automatic personalization of physical rehabilitation exercises in stroke patients," Mathematics, vol. 9, no. 12, p. 1427, 2021.
[18]. J. Davids, N. Lidströmer, and H. Ashrafian, "Artificial Intelligence for Physiotherapy and Rehabilitation," in Artificial Intelligence in Medicine, N. Lidströmer and H. Ashrafian, Eds., Cham: Springer International Publishing, 2022, pp. 1789–1807. doi: 10.1007/978-3-030-64573-1_339.

[19]. F. M. Amiri and A. Khadivar, "A fuzzy expert system for diagnosis and treatment of musculoskeletal disorders in the wrist," Tehnicki Vjesnik/Technical Gazette, vol. 24, 2017, Accessed: Sep. 20, 2024. [Online]. Available: https://www.researchgate.net/profile/Ameneh-Khadivar/publication/317216186_A_fuzzy_expert_system_for_diagnosis_and_treatment_of_musculoskeletal_disorders_in_wrist/links/62dfea639d410c5ff3633309/A-fuzzy-expert-system-for-diagnosis-and-treatment-of-musculoskeletal-disorders-in-wrist.pdf

[20]. F. J. Reis, M. B. L. de Carvalho, G. de Assis Neves, L. C. Nogueira, and N. Meziat-Filho, "Machine Learning Methods in Physical Therapy: A Scoping Review of Applications in Clinical Context," Musculoskeletal Science and Practice, p. 103184, 2024.

[21]. W. Wei, C. McElroy, and S. Dey, "Towards on-demand virtual physical therapist: Machine learning-based patient action understanding, assessment and task recommendation," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 9, pp. 1824–1835, 2019.

[22]. C. Tack, "Artificial intelligence and machine learning| applications in musculoskeletal physiotherapy," Musculoskeletal Science and Practice, vol. 39, pp. 164–169, 2019.

[23]. M. M. Engelhard, "Adaptive Algorithms for Personalized Health Monitoring," PhD Thesis, University of Virginia, 2016. Accessed: Sep. 02, 2024. [Online]. Available: https://scholar.archive.org/work/5gvb3kyqe5b7lhka4kv7j27p5e/access/wayback/https://libraetd.lib.virginia.edu/downloads/dv13zt259?filename=1_Engelhard_Matthew_2016_PHD.pdf

[24]. K. Kifayat, P. Fergus, S. Cooper, and M. Merabti, "Body area networks for movement analysis in physiotherapy treatments," in 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, IEEE, 2010, pp. 866–872. Accessed: Sep. 20, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5480793/

[25]. A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV (version 2.2)," PhysioNet, 2023. doi: 10.13026/6mm1-ek67. [Online]. Available: https://physionet.org/content/mimiciv/2.2/

[26]. A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: An Ensemble Classifier," in International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018, vol. 26, J. Hemanth, X. Fernando, P. Lafata, and Z. Baig, Eds., in Lecture Notes on Data Engineering and Communications Technologies, vol. 26. , Cham: Springer International Publishing, 2019, pp. 758–763. doi: 10.1007/978-3-030-03146-6_86.

[27]. J. Liang, "Confusion matrix: Machine learning," POGIL Activity Clearinghouse, vol. 3, no. 4, 2022, Accessed: Sep. 26, 2024. [Online]. Available: https://pac.pogil.org/index.php/pac/article/view/304

[28]. A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern recognition, vol. 30, no. 7, pp. 1145–1159, 1997.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Blockchain-Based Associate Professor Evaluation System: Transparent Academic Career Management with Ethereum Smart Contracts and IPFS

Kemal Akkanat
Computer Engineering
Trakya University
Edirne, Turkey
kemalakkanat@trakya.edu.tr
ORCID:0000-0002-6448-7378

Muhittin Tanoba
Computer Engineering
Trakya University
Edirne, Turkey
muhittintanoba@trakya.edu.tr
ORCID:0009-0002-6641-759X

Abstract

This paper presents a blockchain-based evaluation system built on Ethereum smart contracts and the InterPlanetary File System (IPFS) to address the structural vulnerabilities of the associate professorship evaluation process in Turkey. The current system suffers from four problems: unverifiable jury selection, risk to document integrity, opaque appeal mechanisms, and a lack of accountability. The proposed framework addresses all four. The system provides on-chain deterministic jury assignment, IPFS-backed hybrid document storage, automated rule-based scoring using a simplified weighting model abstracted from the UAK Associate Professorship Application Criteria, and a transparent state machine governing the full application lifecycle: Pending, JuryAssigned, Accepted, Rejected, Appealed, and Canceled. Implemented on the Ethereum Sepolia testnet using the Truffle v5.11.5 development framework, the system is evaluated through gas cost analysis, comparative system benchmarking, and a security assessment. The evaluation reports gas costs for all core contract functions and a functional comparison against the current process and related blockchain systems. Off-chain identity verification for data entry is included in future development plans.

Keywords: blockchain, smart contract, Ethereum, IPFS, associate professorship, academic evaluation, transparency

I.INTRODUCTION

Academic career advancement in higher education systems rests on formal evaluation processes that weigh research output, teaching experience, and scholarly impact. In Turkey, the associate professorship (doçent) evaluation is a critical milestone in academic career progression, marking the transition from assistant professor to associate professor. The process is coordinated by the Higher Education Council (YÖK) and consists of document submission, jury assignment, multi-criteria scoring, and an appeal mechanism [1]. Despite its institutional importance, the current system carries structural vulnerabilities that undermine stakeholder confidence. Centralized document repositories create a single point of failure, leaving data susceptible to loss or unauthorized modification. Jury selection procedures are not grounded in externally verifiable criteria, which makes the process opaque to candidates and independent observers. Evaluation outcomes are communicated with limited justification, and appeal mechanisms do not offer adequate transparency into the review process. These problems echo broader concerns raised in the blockchain literature about transparency and accountability in centralized systems [2], and in blockchain-in-education research about the gap between credential verification and procedural fairness [3]. Blockchain technology offers an alternative architecture for addressing these structural issues. The defining properties of blockchain (decentralized consensus, cryptographic immutability, and programmable smart contracts) enable trust among multiple stakeholders without relying on any central

authority [4], [5]. Prior research has demonstrated blockchain's effectiveness in academic credential verification [6], [7], multi-actor evaluation and grant management [8], and academic research integrity monitoring [9]. Existing solutions, however, address document issuance and single-institution verification rather than the multi-actor evaluation workflow examined here.

This paper introduces a blockchain-based associate professorship evaluation system built on Ethereum smart contracts and the InterPlanetary File System (IPFS). The proposed framework establishes an auditable record of the evaluation process by recording every step (application submission, jury assignment, voting, and appeal) on an immutable ledger. The principal contributions of this work are listed below:

- On-chain jury assignment: A contract function that selects three jurors from a managed pool using on-chain pseudorandomness, so that every assignment is deterministically reproducible and can later be audited.
- IPFS-backed document integrity protocol: A hybrid storage model in which the full content of credential documents is stored on IPFS. At the same time, only cryptographic hash values are anchored to the blockchain [10].
- On-chain automated scoring model: A calculateScore() function that deterministically computes weighted scores from a parameterizable weighting model abstracted from the UAK Associate Professorship Application Criteria [1] at the time of submission, removing discretionary weighting from the scoring step.
- Transparent state machine and appeal workflow: A formal model that implements the full application lifecycle (Pending, JuryAssigned, Accepted, Rejected, Appealed, Canceled) entirely through smart contract state transitions, supplemented by an on-chain appeal recording mechanism that extends the same procedural guarantees to the secondary independent review.

II.RELATED WORK

Blockchain-based academic credentialing began with Blockcerts [11], which anchors cryptographically signed diplomas to the Bitcoin chain but is constrained by public-chain fees and throughput [2], [4]. EduCTX [12] moves credit transfer onto a consortium blockchain built on the open-source Ark platform and addresses cross-institutional interoperability, while Cerberus [7] mitigates diploma fraud at the accreditation and degree-verification layer; neither model shows how multiple evaluators reach a collective decision. Dash et al. [6] pair Hyperledger Fabric with IPFS for certificate authentication (300–900 ms latency, ~250–280 TPS), and VeritasChain [9] extends Fabric-based credentialing with research-integrity monitoring and zero-knowledge proofs (reported 182 TPS, 1.2 s verification). Both remain single-decision verification systems with no jury layer. On the governance side, Hjalmarsson et al. [13] demonstrate Ethereum-based e-voting without addressing voter identity verification, and Erturk et al. [14] illustrate the on-chain-hash/off-chain-content pattern [10] in a non-credentialing setting. Closest to this work is the fellowship management system of Gupta et al. [8], which combines eligibility verification, attribute-based encryption, and zero-knowledge proofs at a reported $0.39–$1.31 per operation, but is specific to funding allocation and provides no application scoring, jury pool management, or appeal handling. Across these works, blockchain research has concentrated on single-institution document verification or isolated voting decisions [3], [11–14]. No existing system integrates a multi-actor evaluation workflow, pool-managed jury assignment, weighted scoring against an institutional standard such as the UAK criteria, and an on-chain appeal mechanism in a single contract architecture. The proposed system is designed to close this gap.

III.SYSTEM MODEL

1. System Actors and Roles

Four actors interact with the system, each with authority boundaries encoded in the contract. The Candidate uploads supporting documents to IPFS, submits the resulting content identifier (CID) via createApplication(), and may appeal a rejection by writing the grounds on-chain. The YÖK Administrator (yokAdmin), a single Ethereum address fixed at deployment, maintains the jury pool and triggers assignment. Jury Members evaluate an assigned application and cast one irrevocable vote each. The Auditor/Institution is any external observer who can query application histories, jury assignments, and voting outcomes without prior authorization [2].

2. Layered Architecture

The system comprises three layers. The application layer is a single-page JavaScript application with role-based panels for candidate, jury member, and administrator roles, authenticated via MetaMask and interacting with the contract via Ethers.js v6. The smart contract layer holds all business logic in DocentlikSistemi.sol (Solidity 0.8.20), deployed to the Ethereum Sepolia testnet with Truffle v5.11.5. The storage layer holds document content on IPFS via Pinata, with only the CID persisted on-chain [10].

3. Application Lifecycle (State Machine)

Each application in the system progresses through a six-state state machine enforced by the smart contract (Table 1). Every state transition can be triggered only by the authorized actor when the specified conditions are satisfied; the transition logic is enforced at runtime by require() checks within the contract.

Table 1. Application State Machine

State	Triggering function	Authorized actor	Successor state(s)
Pending	createApplication()	Candidate	JuryAssigned, Canceled
JuryAssigned	assignJury()	YÖK administrator	Accepted, Rejected, Canceled
Accepted	castVote() – 2/3 for	Jury members	Terminal
Rejected	castVote() – 2/3 against	Jury members	Appealed; Terminal if appealCount = MAX_APPEAL
Appealed	appeal()	Candidate	JuryAssigned
Cancelled	cancelApplication()	YÖK administrator	Terminal

Transitions are guarded on-chain: castVote() requires the JuryAssigned state and appeal() requires Rejected, so these constraints hold regardless of front-end behavior [5].

IV. SMART CONTRACT DESIGN

1. Data Structures

The contract is built on three core data structures. The Application struct below consolidates all evaluation data for a given candidate into a single record:

```
struct Application {
address candidate; uint256 score; string ipfsHash; Status status; uint256 adminYears;
```

string rejectionReason; uint256 submissionTime; uint256 pubCount; uint256 citeCount;}

Three mappings hold auxiliary state: applicationJuries (id → three juror addresses), juryDecisions (id → juror → vote, where 0 = not voted, 1 = accept, 2 = reject), and appealCounter (id → number of appeals filed). Vote tallies are not stored; castVote() recomputes them from juryDecisions on each call.

2. Scoring Model

The calculateScore() function deterministically computes a weighted score from the three-parameter model summarized in Table 2, which abstracts the publication, citation, and teaching categories of the UAK Associate Professorship Application Criteria [1]. Declared as pure, the function neither reads nor modifies chain state, so it consumes no gas when queried externally via eth_call; when invoked internally during createApplication(), its cost is included in the gas reported for that function:

The total score is expressed as $P = 8 \cdot n_pub + \lfloor n_cite/2 \rfloor + 5 \cdot n_admin$, where n_pub is the declared publication count, n_cite is the international citation count, and n_admin is the number of administrative or teaching years. The prototype applies a single publication weight; differentiating SCI/SSCI-indexed and nationally indexed publications is left to future work. Solidity has no fixed-point type, so the citation weight is implemented as integer division that truncates toward zero: two citations contribute one point. The contract enforces the UAK weighting deterministically, but does not verify the declared counts themselves, which are supplied by the candidate at submission; automated validation against YOKSIS records is required before the scoring step can be considered trustless (Section VIII).

Table 2. Score Weights Based on UAK Associate Professorship Application Criteria

Criterion	Contract parameter	Weight	Contribution
SCI/SSCI-indexed publication	PubCount	× 8	8·n_pub
International citation	citeCount	÷ 2	⌊n_cite/2⌋
Administrative / teaching year	adminYears	× 5	5·n_admin

A minimum score threshold can be defined as a contract constant; applications falling below that threshold can be automatically rejected during the required check in submitApplication(), thereby directing jury capacity exclusively toward qualified candidates [8].

3. Jury Assignment Mechanism

Jury assignment is the core mechanism that makes the system independent of a central decision body. Algorithm 1 presents the pseudocode for the assignJury() function:

Algorithm 1: assignJury(applicationId)

1. require(status == Pending or Appealed)
2. require(activeJuryCount >= 3)
3. seed ← keccak256(blockhash(block.number-1), applicationId, timestamp)
4. active ← array of active jury addresses
5. for i ← 0..2: swap actives[i] ↔ actives[i + keccak256(seed,i) mod (n−i)]
6. selected[0..2] ← actives[0..2]

7. applicationJuries[id] ← selected; status ← JuryAssigned

The current prototype derives randomness from blockhash(block.number - 1). This makes each assignment deterministically reproducible from public chain data, and therefore auditable, but not verifiably random in the cryptographic sense. Two weaknesses follow. A miner can, in principle, influence the block hash. More practically, because blockhash(block.number - 1) is already known to the caller when the transaction is composed, the administrator can compute the resulting panel in advance and withhold the call until a favorable draw appears. Replacing this layer with Chainlink VRF [15] removes both and is a prerequisite for production deployment.

4. Voting and Consensus Mechanism

Jury members cast their votes independently via the castVote() function. When the third vote is recorded, the contract automatically applies a 2/3 majority rule. The voting mechanism enforces three security properties: the jury decisions mapping prevents a jury member from voting on the same application more than once; votes are irrevocable; and the 2/3 decision threshold is hard-coded in castVote() with no setter exposed, so no one, including the administrator, can alter it.

5. Appeal Process

A rejected candidate can invoke the appeal() function to write their grounds for appeal directly to the chain. The contract records the appeal and updates the status to Appealed. Every appeal step is written to an immutable record, so initial and appeal decisions are equally auditable [6], [9]. The MAX_APPEAL constant, set at contract deployment, prevents an infinite appeal loop from being opened against the same rejection decision. The post-appeal process carries the same procedural guarantees as the initial evaluation: assignJury() is called again, a new panel is drawn from the same active pool with a fresh seed, and the voting process restarts from scratch. The current implementation does not prevent the original panel from being reselected; enforcing panel disjointness is left to future work.

V.DOCUMENT MANAGEMENT AND IPFS INTEGRATION

Application documents such as CVs, publication lists, diplomas, and project reports range from a few hundred kilobytes to several megabytes. Persisting such content on-chain is impractical: under the EVM storage schedule, a newly written 32-byte slot costs 20,000 gas, about 640,000 gas per kilobyte, and blocks are additionally bounded by the block gas limit [5]. The system therefore adopts the hybrid pattern used in comparable work [6], [9]: content is pinned to IPFS through Pinata, and only the resulting CID is written to Application.ipfsHash by createApplication(), holding on-chain state at 1–2 KB per record regardless of document size [10]. Because an IPFS CID is the root hash of a Merkle DAG built over SHA-256 chunk hashes, any single-byte modification yields a different CID; assigned jurors read documents directly from IPFS, and any third party can verify post-submission integrity against the on-chain CID without an intermediary. By comparison, an off-chain server with a URL pointer costs only ≈40,000 gas but offers no integrity guarantee, since the target content remains mutable. Pinata provides primary pinning, and Filecoin archiving is planned as a second tier. The on-chain CIDs remain valid identifiers even in the event of a storage-layer failure, but their retrievability depends on at least one node continuing to pin the content. Because the CID is public, any observer can also retrieve the underlying documents. The prototype, therefore, assumes that submitted material is disclosable; handling personal data under KVKK and GDPR, including the conflict between ledger immutability and the right to erasure, would

require encrypting content before pinning and anchoring, with decryption entitlements on-chain only, as in [8]. This is outside the present scope (Section VIII).

VI.IMPLEMENTATION AND EVALUATION

1. Implementation Environment

The proposed system was developed for the Ethereum Sepolia testnet, then compiled and deployed using the Truffle v5.11.5 development framework. The smart contract is written in Solidity 0.8.20; the front-end is built with JavaScript and the Ethers.js v6 library. Wallet connectivity is provided via the MetaMask extension, and document storage is handled through the Pinata IPFS pinning service [10].

2. Gas Cost Analysis

Gas consumption on the Ethereum Virtual Machine (EVM) is deterministic for a given EVM version and storage state, so local measurements under Ganache agree with Sepolia values for the operations reported here; the figures in Table 3 were taken from the Sepolia deployment identified in Section VI-E. The US-dollar equivalents in Table 3 are derived from two explicitly stated parameters: a gas price of 1 Gwei and an ETH/USD rate of 1,940, the spot rate on 26 July 2026. The 1 Gwei assumption is deliberately conservative, so the dollar figures should be read as upper bounds rather than typical costs. Table 3 summarizes both quantities for the core functions. assignJury() is the only function whose cost scales with system size, since it enumerates the active jury pool in memory before sampling (Table 3, note b); extrapolated to the thousand-reviewer pool a national board would maintain, it approaches the block gas limit. Indexing the pool rather than enumerating it is therefore a prerequisite for production deployment, and is the principal scaling constraint of the present design.

Table 3. Smart Contract Function Gas Costs

Function	Gas used	Cost (USD)[a]
createApplication()	196,852	0.382
assignJury()[b]	186,236	0.361
castVote()	63,178	0.123
appeal()	81,695	0.158
calculateScore()	0 (external call)	0.000

[a] A dollar equivalent assumes a gas price of 1 Gwei and ETH/USD = 1,940 (26 July 2026); at the sub-1 Gwei averages typical of 2026 mainnet conditions, the corresponding costs are proportionally lower. calculateScore() is declared pure; it consumes no gas when queried externally via eth_call, while its cost during on-chain submission is contained in the createApplication() figure. [b] Measured for a pool of 10 jury members. The cost rises to 437,324 gas for 50 members and 751,194 gas for 100 members, an increase of approximately 6,277 gas per additional pool member, confirming linear scaling with pool size.

The calculateScore() function incurs no gas since it is declared as pure [5], making frequent score queries cost-free. Under the assumptions above, per-transaction costs fall in the $0.12–$0.38 range, below the $0.39–$1.31 reported for the comparable smart contract system in [8]. The comparison is indicative rather than exact, since the figures in [8] were priced under different market conditions; the underlying gas measurements, being independent of token price, provide the more durable point of comparison.

3. System Comparison

Table 4 presents a comparison of the proposed system with the existing YÖK-centric evaluation process and the most closely related works across six functional criteria. The comparison methodology follows a similar structure to related permissioned-blockchain credential studies [6]. Because the reviewed systems were not designed for the same task, the comparison is qualitative. Each cell records whether a capability is claimed in the corresponding publication, not how well it performs.

Table 4. Functional Comparison Across Six Criteria

System	C1	C2	C3	C4	C5	C6
Current YÖK process	M	N	N	N	P	P
Blockcerts [11]	N	N	P	N	N	Y
EduCTX [12]	N	N	P	N	N	Y
Cerberus [7]	N	N	N	N	N	Y
Dash et al. [6]	N	N	Y	N	N	Y
VeritasChain [9]	N	N	P	N	N	Y
Gupta et al. [8]	P	N	N	N	N	Y
Proposed system	Y	Y	Y	Y	Y	Y

C1 multi-actor evaluation workflow; C2 verifiable jury assignment; C3 hybrid document integrity (on-chain hash, off-chain content); C4 automated scoring against an institutional standard; C5 on-chain appeal mechanism; C6 third-party auditability. Y = supported, P = partially supported, N = not supported, M = supported but performed manually and off-record. Cell values characterize the functional scope that each system reports in its own publication, as summarized in Section II; they are not performance measurements.

4. User Interface

The system provides three separate role-based panels for the candidate, jury member, and YÖK administrator. Following MetaMask authentication, each panel is accessible only to the Ethereum address associated with the relevant role; interface-layer role restrictions are independently validated by require() checks at the smart contract level [5]. As a worked example of the scoring model end-to-end, an application declaring two publications, three international citations, and five administrative years returns $2 \cdot 8 + \lfloor 3/2 \rfloor + 5 \cdot 5 = 42$ points, which is the value written on-chain and displayed in the panel (Figure 1).

The prototype's front-end was implemented in Turkish for the pilot deployment shown in Figure 1 (e.g., "Aday Paneli", "Yeni Başvuru Oluştur", "Jüri Atandı"); these interface labels map directly onto the English contract functions and fields used throughout this paper (createApplication(), pubCount, citeCount, adminYears) and do not affect on-chain behavior.

The Jury Member panel surfaces only the IPFS-linked application documents and the castVote() control for applications currently assigned to that address; no scoring or administrative controls are exposed at this interface, mirroring the on-chain role restriction enforced by require(). The YÖK Administrator panel is limited to jury-pool management and the triggering of assignJury(). It does not expose any control that can alter a submitted score or vote once cast, consistent with the immutability guarantees described in Section VII. Each panel queries the contract directly through Ethers.js v6 rather than through an intermediary backend, so the data displayed to every role is read from the same on-chain state that any external observer can independently verify via a public block explorer such as Etherscan.

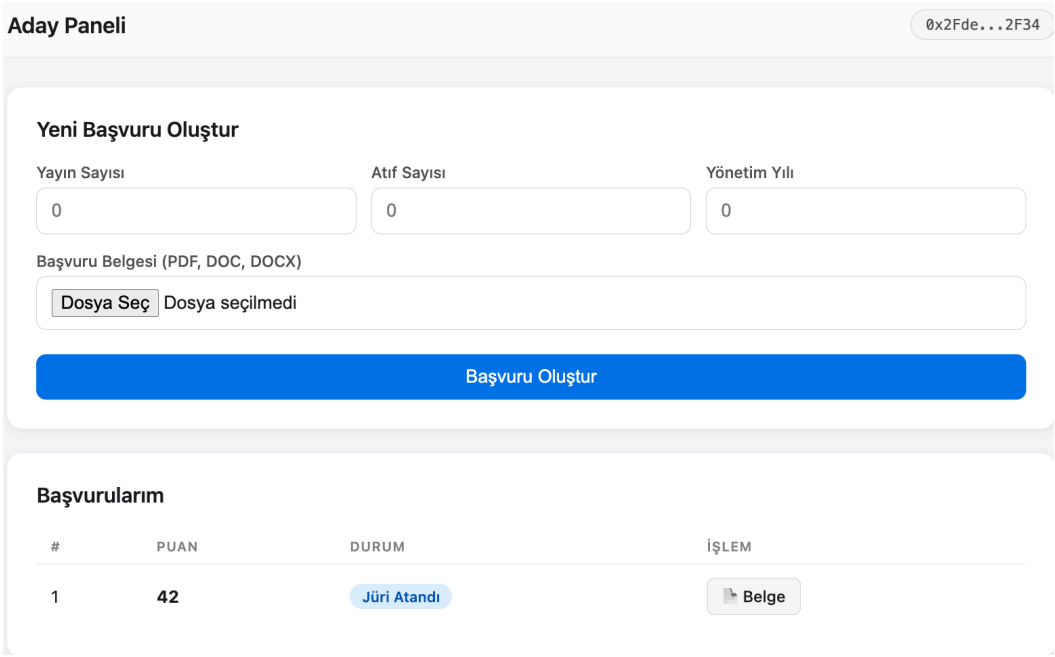


Figure 1. Candidate application interface and IPFS document upload.

5. Reproducibility and Artifact Availability

The gas figures in Table 3 were obtained with eth-gas-reporter against a local EVM instance of DocentlikSistemi.sol. These figures carry over to any EVM-compatible network, since gas consumption is deterministic for identical bytecode and inputs; the contract was also deployed to the Sepolia testnet, where the deployment cost of 2,040,811 gas matches the local measurement exactly. The contract source, migration scripts, and the 49-case Truffle test suite are available in the source repository, and the deployed contract, together with its complete transaction history, is publicly queryable on Etherscan. Deployed contract: 0xD3936E8639b4862F4DB2a33Cab69b50B38c1c9eE (Sepolia). Source repository: github.com/MuhittinTanoba/doc.

VII. SECURITY ANALYSIS

The analysis framework draws on the threat-countermeasure structure used in related permissioned-blockchain credential work [6] and the Ethereum smart contract vulnerability literature [16].

1. Tamper Resistance and Access Control

Modifying data committed to the Ethereum blockchain is computationally infeasible. Once an application record is created and its fields are written to the applications[id] mapping, they are protected by an immutability guarantee [5]. Document contents are guaranteed by IPFS content-addressed structures, as described in Section V [10].

Every critical function in the contract verifies the message. sender address through either the onlyAdmin modifier or a require() check. The assignJury() function can only be called from the YokAdmin address set at

deployment; castVote() is restricted to one of the three addresses registered in the applicationJuries[id] array. The EVM evaluates these constraints at call time, so unauthorized calls revert before any state is written.

2. Double Voting and Replay Resistance

The jury decisions [id][msg.sender] mapping deterministically ensures that each jury member may cast only one vote per application. When a member calls castVote() a second time, the require(juryDecisions[id][msg.sender] == 0) condition fails, and the transaction is reverted.

Replay resistance follows from record identity: each application is timestamped with a block. timestamp and applicationId increments monotonically with every new record, so an earlier submission cannot be resubmitted as a new one. Repeated appeals against the same rejection decision are bounded by the MAX_APPEAL check described in Section IV-E.

3. Insider Threat Mitigation

The YÖK administrator can assign jury members but cannot alter the voting outcome, modify a candidate's score, or delete an appeal rationale, as these data become immutable once written to the contract. All transactions are publicly queryable via Etherscan. A multi-signature scheme for critical administrative functions is left to future work.

4. Smart Contract Vulnerabilities

Reentrancy attacks, classified by Atzei et al. as the most critical vulnerability category in Ethereum smart contracts [16], are structurally eliminated in the proposed contract by the absence of any function that transfers ETH to an external address. Before deployment, 49 unit tests were executed under the Truffle test suite, covering key boundary conditions and invalid state transitions. Independent formal verification and a smart contract audit are planned steps before any production release.

VIII. CONCLUSION AND FUTURE WORK

This paper presented a blockchain-based framework for evaluating associate professorships that addresses four structural weaknesses in the current process: unverifiable jury selection, document integrity risks, opaque appeals, and weak accountability. These are addressed through on-chain deterministic jury assignment, IPFS-anchored document integrity, automated UAK-based scoring, and a six-state transparent lifecycle with an on-chain appeal path. Deployed on the Sepolia testnet, the system executes its core functions at 63,178–196,852 gas, corresponding to $0.12–$0.38 under the pricing assumptions stated in Section VI-B, below the $0.39–$1.31 range reported for the closest comparable system [8]. Several limitations remain. Blockhash-based pseudorandomness is theoretically susceptible to miner influence and, more practically, to pre-computation by the caller; it will be replaced by Chainlink VRF [15]. The declared publication and citation counts are not validated against an external register: the system is not yet integrated with YOKSIS or the e-Government identity infrastructure. Hence, data entry still requires an off-chain verification step, which constrains the trustless model. Document CIDs are written in the clear, so confidentiality of candidate material and compliance with KVKK and GDPR require an encryption layer that the prototype does not implement. Finally, assignJury() scales linearly with the size of the jury pool, and the evaluation was confined to a testnet; therefore, institutional-scale deployment requires further validation. Future work will additionally explore ZK-SNARK-based confidential voting [17] and a DAO-based governance model [18].

REFERENCES

[1]. Interuniversity Board (UAK), "Associate Professorship Application Requirements and Scoring Tables," 2025. Available: https://www.uak.gov.tr/page/docentlik-basvuru-sartlari-kLPHX

[2]. Z. Zheng, S. Xie, H.-N. Dai, X. Chen, and H. Wang, "Blockchain challenges and opportunities: A survey," Int. J. Web Grid Serv., vol. 14, no. 4, pp. 352–375, 2018. DOI: 10.1504/IJWGS.2018.095647

[3]. P. Ocheja et al., "Blockchain in education: A systematic review and practical case studies," IEEE Access, vol. 10, pp. 99525–99540, 2022. DOI: 10.1109/ACCESS.2022.3206791

[4]. S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008. Available: https://bitcoin.org/bitcoin.pdf

[5]. G. Wood, "Ethereum: A secure decentralized generalized transaction ledger," Ethereum Yellow Paper, 2014. Available: https://ethereum.github.io/yellowpaper/paper.pdf

[6]. S. P. Dash, A. K. Jena, and D. K. Murala, "A Hyperledger-based secure framework for academic certificate authentication using blockchain," Int. J. Saf. Secur. Eng., vol. 15, no. 6, pp. 1185–1195, 2025. DOI: 10.18280/ijsse.150610

[7]. A. Tariq, H. B. Haq, and S. T. Ali, "Cerberus: A blockchain-based accreditation and degree verification system," IEEE Trans. Comput. Social Syst., vol. 10, no. 4, pp. 1503–1514, 2023. DOI: 10.1109/TCSS.2022.3188453

[8]. B. K. Gupta, A. Gupta, J. Alam, and S. Maity, "Blockchain-based fellowship management system with zero-knowledge credential verification," Proc. 13th SMART, IEEE, 2024. DOI: 10.1109/SMART63812.2024.10882489

[9]. S. Tahmina and Md. A.-A. Bhuiyan, "VeritasChain: A decentralized blockchain for academic and research integrity," Proc. 28th ICCIT, IEEE, 2025. DOI: 10.1109/ICCIT68739.2025.11490198

[10]. J. Benet, "IPFS — Content-addressed, versioned, P2P file system," arXiv:1407.3561, 2014. DOI: 10.48550/arXiv.1407.3561

[11]. MIT Media Lab & Learning Machine, "Blockcerts — An open infrastructure for academic credentials on the blockchain," 2016. Available: https://www.blockcerts.org

[12]. M. Turkanovic et al., "EduCTX: A blockchain-based higher education credit platform," IEEE Access, vol. 6, pp. 5112–5127, 2018. DOI: 10.1109/ACCESS.2018.2789929

[13]. F. T. Hjalmarsson et al., "Blockchain-based e-voting system," Proc. 2018 IEEE 11th Int. Conf. Cloud Computing (CLOUD), pp. 983–986. DOI: 10.1109/CLOUD.2018.00151

[14]. E. Erturk et al., "NFT-based fundraising system for preserving cultural heritage: Heirloom," Proc. 6th UBMK, IEEE, 2021. DOI: 10.1109/UBMK52708.2021.9559006

[15]. Chainlink Labs, "Chainlink VRF: Verifiable random function for smart contracts," 2020. Available: https://docs.chain.link/vrf

[16]. N. Atzei, M. Bartoletti, and T. Cimoli, "A survey of attacks on Ethereum smart contracts (SoK)," Proc. 6th Int. Conf. Principles of Security and Trust (POST), LNCS vol. 10204, pp. 164–186, Springer, 2017. DOI: 10.1007/978-3-662-54455-6_8

[17]. J. Groth, "On the size of pairing-based non-interactive arguments," Proc. EUROCRYPT 2016, LNCS vol. 9666, pp. 305–326, Springer, 2016. DOI: 10.1007/978-3-662-49896-5_11

[18]. V. Buterin, "Ethereum: A next-generation smart contract and decentralized application platform," Whitepaper, 2014. Available: https://ethereum.org/en/whitepaper/

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Region-Ablation Study of MRI Texture Features for Trigeminal Neuralgia Classification

Önder Polat
Department of Electrical and Electronics Engineering
Gaziantep Islam Science and Technology University
Gaziantep, Türkiye
onder.polat@gibtu.edu.tr
ORCID: 0000-0002-0439-3994

Habip Eser Akkaya
Department of Radiology
Ankara Training and Research Hospital
Ankara, Türkiye
habipeser.akkaya@saglik.gov.tr
ORCID: 0000-0002-8447-3627

Zeynep Ziroğlu
Department of Neurology
Ankara Training and Research Hospital
Ankara, Türkiye
drzeynepkuzu@hotmail.com
ORCID: 0000-0002-2859-5072

Hafize Nalan Güneş
Department of Neurology
Ankara Training and Research Hospital
Ankara, Türkiye
gokcegunes@yahoo.com
ORCID: 0000-0002-9719-2330

Abstract
In medical image classification, features are usually computed inside a region of interest (ROI) drawn by an expert. This assumes that the information needed to separate the classes lies within the annotated structure. We tested this assumption directly on a brain Magnetic Resonance Imaging (MRI) classification task, treating the region rather than the feature set or the classifier as the experimental variable. From 80 subjects with expert annotations of a small nerve structure, we extracted the same intensity and texture descriptors from seven spatially distinct regions: the annotated interior; concentric shells of 4, 6, 8, and 10 mm around it; a 40 mm enclosing box; and a fixed-geometry spherical shell control. All regions were evaluated with an identical pipeline and identical repeated cross-validation. Interior features showed an Area Under Curve (AUC) of 0.555 ± 0.136. Shell regions achieved an AUC of 0.743–0.800, whereas the enclosing box achieved an AUC of 0.545 ± 0.114. To test whether shell performance was inherited from the geometry of the annotation, we repeated the experiment with a fixed-radius spherical shell placed at the annotation centroid, identical in shape for every subject; this reached 0.633 ± 0.137, above chance, but noticeably below the mask-derived shells. The discriminative content, therefore, lies in a narrow band just outside the annotation. Annotation location alone carries some of this signal, but most of it depends on proximity to the annotation boundary itself, and the effect is destroyed by enlarging the region. This effect reflects genuine image content rather than annotation shape. In this task, the location of the features mattered more than which features or which classifier were used.

Keywords: medical image analysis, trigeminal neuralgia, region of interest, texture features, ablation study, MRI, feature extraction

I. INTRODUCTION

Trigeminal neuralgia (TN) is a chronic facial pain disorder. Patients experience sudden, brief, extremely severe electric-shock-like pain on one side of the face, typically triggered by ordinary actions such as chewing, speaking, or touching the skin. Diagnosis rests primarily on clinical criteria [1], [2]. The relevant anatomy is the trigeminal nerve, a structure only a few millimeters across that runs from the brainstem through a fluid-filled space toward the skull base. The prevailing explanation for the disorder is mechanical: a nearby blood vessel

presses on the nerve, gradually damaging its insulating sheath. This mechanical injury is associated with measurable nerve volume loss, which has also been linked to outcomes after surgical decompression [3]. Because the fluid surrounding the nerve appears bright on heavily T2-weighted Magnetic Resonance Imaging (MRI) while the nerve and vessels appear dark, this region can be imaged at sub-millimeter resolution, and high-resolution T2-weighted MRI is the established imaging protocol when the diagnosis is suspected [4]. From an engineering standpoint, however, this is an unusually difficult classification problem, for three reasons. First, the target structure is very small, on the order of a few hundred cubic millimeters, so the number of voxels available per subject is limited and partial-volume effects are substantial. Second, and more importantly, the imaging finding is not decisive. Vascular contact between the nerve and a vessel is common even in individuals without the condition, whereas findings severe enough to distort or compress the nerve are comparatively rare [5]. Distinguishing between these degrees of contact from a radiological read alone is subjective, which is precisely the situation in which automated quantitative analysis is attractive. Third, cohorts are small. TN is uncommon, and datasets with expert annotations typically number in the tens to low hundreds, a regime in which conventional handcrafted predictors remain competitive with, and easier to interpret than, end-to-end deep learning models.

II. RELATED WORK

Extracting a large vector of quantitative descriptors from a region of interest (ROI) and mining it for diagnostic or prognostic signal is known as radiomics [6], [7]. It was originally developed for oncologic imaging but has since been applied broadly across anatomical targets, including small non-tumoral structures such as peripheral nerves. In this study, we also utilized radiomics as the general framework. Several groups have applied this framework specifically to TN. Wang et al. [8] extracted a large radiomic feature vector from MRI in a cohort of 350 patients and 100 controls, reduced dimensionality with principal component analysis followed by LASSO, and compared regularized logistic regression, support vector machines, and boosting, reporting high accuracy with prospective validation on an additional small set. A study published in [9] combined automatic segmentation of the nerve using a U-Net with the extraction of 216 descriptors, and related them to symptom presence in patients treated with radiosurgery. A more recent nomogram study [10] extended the region definition by extracting features from two adjacent anatomical compartments, along with hand-measured anatomical parameters, and reported that the combined model outperformed image descriptors alone. Deep learning approaches have also been reported. A 2025 study using convolutional networks on plain radiographs [11] achieved 87.2% accuracy on an internal test set but 71.0% on external data, a 16-point drop that illustrates how fragile reported figures in this area can be. A separate line of work [12] segmented both the nerve and the adjacent vessels and, notably, cropped a fixed cube around the nerve centroid rather than masking to the nerve itself.

III. MOTIVATION

Across most of this literature, one design choice is shared and rarely examined: descriptors are computed inside the expert annotation. This encodes the assumption that the class-discriminative information is contained within the annotated structure. For a condition defined by the relationship between a structure and its immediate surroundings, that assumption is not obviously correct, and the fixed-cube cropping used in [12] implicitly departs from it, though without a direct comparison. We therefore keep the feature set, the classifier, and the evaluation protocol fixed, and vary only the region from which features are computed. We report an ablation across four regions derived from the same annotation, and a control experiment that separates image content from annotation geometry. The contribution of this study is an empirical result on where the signal lives and a control procedure we recommend for any study using annotation-derived context regions.

IV. METHOD

1. Dataset

Ninety-two subjects (45 patients, 47 controls) were retrospectively collected at the Ankara Training and Research Hospital. This retrospective study was approved by the Ankara Training and Research Hospital Ethics Committee (approval no. E-23-1363, decision dated 07.09.2023). For every subject, an experienced neuroradiologist manually delineated the region containing the course of the trigeminal nerve on the axial volume. Annotations were bilateral in most subjects. Acquisition was heterogeneous. Twelve patients, and no controls, had been imaged with a 3D steady-state sequence (constructive interference in steady state, CISS) rather than T2-weighted imaging. This imbalance arises naturally in retrospective archives, since that sequence is typically ordered only when the condition is clinically suspected. Because such an imbalance could be learned by the model rather than the diagnostic label, these 12 patients were excluded, and all subsequent experiments were restricted to the sequence-homogeneous T2-weighted subcohort (33 patients, 47 controls, n = 80). Demographic characteristics of the patient and control groups are summarized in Table 1.

2. Region Definitions

Four regions were derived per subject from the same expert annotation, denoted M (Figure 1):

- Interior: the annotated region M itself.
- Shell(r): a thin band immediately surrounding M, obtained by expanding M outward by r millimeters and then subtracting the original region, for $r \in \{4, 6, 8, 10\}$ mm.
- Box: a 40 mm cube centered on the midpoint of M.
- Fixed spherical shell (control): all voxels lying between 3 and 7 mm from the midpoint of M.

The fixed spherical shell is the key control. Unlike the other three regions, its shape and size are identical for every subject; the only thing it inherits from the annotation is its location. If the shell effect were caused only by differences in how the masks were drawn, this fixed-shape region would show no effect. It does show the effect, so the result comes from the image, not from the mask shape. For subjects with bilateral annotations, each side was treated as an independent connected component sharing the same diagnostic label; each component was assigned its own centroid when constructing the box and fixed spherical shell regions.

Table 1. Demographic Characteristics of the Study Cohort

Characteristics	Patients	Controls	p
Age, years (mean ± SD)	56.98 ± 15.65	54.46 ± 14.53	0.402
Sex, female (%)	51	67.3	0.096

3. Feature Extraction

Features were computed using PyRadiomics [13] following the Image Biomarker Standardization Initiative (IBSI) conventions [14]. Volumes were resampled to 0.6 mm isotropic spacing with B-spline interpolation, intensity-normalized, and discretized with a fixed bin width of 25. Six descriptor families were enabled: first-order statistics, Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray

Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), and Neighborhood Gray Tone Difference Matrix (NGTDM), yielding 93 features per region.

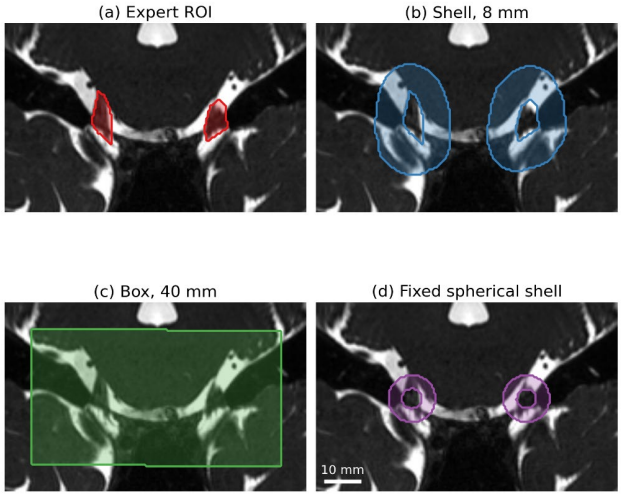


Figure 1. The four regions evaluated in this study are shown on a representative axial slice from a subject with bilateral annotation. (a) the expert-drawn ROI; (b) the shell region, illustrated at r = 8 mm; (c) a 40 mm box; (d) the fixed spherical shell control (radius 3–7 mm from each nerve's own centroid).

Shape descriptors were excluded from all region models. Shape features characterize the geometry of the region rather than the image content within it, and are not comparable across regions with differing construction; for regions 3 and 4, they are constant by design. For reference, the full 107-feature set, including shape, computed on the interior region, gives an AUC of 0.817 ± 0.092, which is noticeably higher than the texture-only result below, since shape descriptors partly encode annotation size and could reflect either genuine anatomical difference or annotator bias, a question we do not attempt to resolve here. From a computational standpoint, each of the six descriptor families summarizes the region through a different mathematical construction. First-order statistics operate directly on the voxel intensity histogram, independent of spatial arrangement. The remaining five families are texture matrices that instead encode spatial relationships between voxel intensities: GLCM tabulates how often pairs of gray levels co-occur at a fixed offset and direction, then averages the result over multiple directions to obtain rotation-invariant descriptors; GLRLM and GLSZM measure the length or size of contiguous runs and zones of constant gray level, capturing coarseness versus fineness of the texture; GLDM quantifies local gray-level similarity within a fixed neighborhood; and NGTDM measures the difference between a voxel's intensity and the average intensity of its neighbors, capturing local contrast. Because these matrices are computed independently within each region mask, their values are sensitive to the number of voxels available: very small or thin regions yield sparser co-occurrence and run-length matrices, which is why a fixed bin width was used for intensity discretization, as recommended by IBSI [14] to keep texture matrices comparable across regions of different sizes.

4. Evaluation

Every region was evaluated with an identical pipeline, implemented with scikit-learn [15]. Four steps were applied in sequence. First, features with near-zero variance across subjects were removed, since a feature that barely changes from one subject to another carries little information for classification. Second, the remaining

features were standardized to zero mean and unit variance, so that no feature would dominate the model simply because it was measured on a larger numerical scale. Third, univariate feature selection was used to retain only the 20 features most strongly associated with the diagnostic label, out of the 93 available; this step reduces the risk of overfitting, which is a particular concern with a sample size of this magnitude. Finally, an L2-regularised logistic regression classifier was trained on these 20 features. Logistic regression was chosen for its simplicity, interpretability, and stable behavior on small datasets, and the L2 penalty discourages the model from assigning excessively large weights to any single feature. All four steps were fitted using only the training fold in each cross-validation split. This is because if feature selection were performed on the full dataset before cross-validation, information from the test fold would leak into the selection process, artificially inflating the reported performance. Performance was estimated using repeated stratified 5-fold cross-validation with 20 repetitions, giving 100 train-test splits per region. Each subject contributed exactly one feature vector per region, computed jointly across both nerves in bilateral cases; cross-validation was therefore inherently performed at the subject level, with no risk of the same subject appearing in both the training and test set within a fold. Results are reported as the mean AUC across these 100 splits, together with the standard deviation and the 2.5th-97.5th percentile range, which shows where the central 95% of the 100 observed AUC values actually fell. Because the 100 splits are not fully independent, each subject contributes to multiple repetitions, the reported percentile range should be interpreted as a description of observed variability across splits rather than a formal confidence interval.

V. RESULTS

Table 2 summarizes classification performance across all seven regions. Four observations were derived from these results (Figure 2). First, features within the annotated region contained no diagnostic information. All 93 descriptors computed inside the expert-drawn boundary performed at the level of random guessing, with an average AUC close to 0.5. In practical terms, texture and intensity patterns measured strictly within the nerve itself did not help separate patients from controls. Second, features from just outside the annotated region did carry useful information. When the same 93 descriptors were computed within a thin band surrounding the annotation, performance rose sharply, reaching 0.743-0.800 depending on how far outward the band extended (4, 6, 8, or 10 mm). Every one of these four distances outperformed both the interior region and the large box. This difference was statistically significant; a Wilcoxon signed-rank test comparing paired fold-level AUC values between the interior region and the 8 mm shell (the best-performing shell radius) confirmed the gap was not attributable to chance ($p < 0.001$). The differences among the four distances themselves were small compared to the normal variability seen from one cross-validation fold to the next. The more important finding is that performance remained consistently high across the entire 4-10 mm range, rather than depending on selecting exactly the right distance. Third, this effect is only partially explained by the annotation's location. A natural concern is that the band surrounding the annotation might only be picking up differences in annotation shape or size between the two groups, rather than any real property of the underlying image. To test this, we repeated the analysis using a region with a completely fixed shape and size, a spherical shell that was identical for every subject, and whose only connection to the original annotation was its center point. This control achieved an AUC of 0.633, which is shown as a reference line in Figure 2. This is clearly above chance, showing that location alone carries some signal, but it remained noticeably lower than the mask-derived shells (0.743–0.800). This gap indicates that shell performance depends not only on being near the annotation, but on being adjacent to its actual boundary. As a further check, we measured how strongly the 4 mm shell features were correlated with the volume of the original annotation; this correlation was weak (median absolute correlation of 0.158), which rules out a purely geometric artifact even though location alone does not account for the full effect. AUC was used as the primary metric because it is threshold-independent and robust to moderate class imbalance. It represents the probability that a randomly selected positive instance receives a higher predicted score than a randomly selected negative instance [16].

Table 2. Classification Performance by Region

Region	AUC (mean ± SD)	[2.5–97.5%]
Interior (expert annotation)	0.555 ± 0.136	0.293–0.830
Shell, 4mm	0.743 ± 0.106	0.496–0.927
Shell, 6mm	0.778 ± 0.101	0.543–0.960
Shell, 8mm	0.800 ± 0.096	0.609–0.969
Shell, 10mm	0.746 ± 0.105	0.553–0.942
Box, 40mm	0.545 ± 0.114	0.348–0.724
Fixed Spherical Shell (control)	0.633 ± 0.137	0.406–0.889

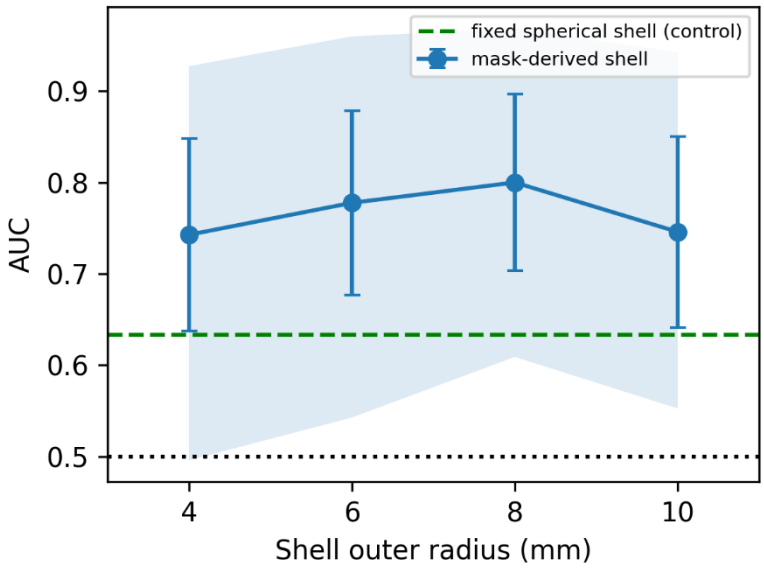


Figure 2. AUC as a function of shell radius (4–10 mm), with the fixed spherical shell control (AUC = 0.633) shown as a horizontal reference line. The control lies below all four shell points, indicating that proximity to the true annotation boundary contributes signal beyond location alone. Error bars show mean ± one standard deviation; the shaded band shows the 2.5th–97.5th percentile range across the 100 cross-validation splits.

VI. DISCUSSION

The main finding is that the choice of features mattered more than the choice of classifier. The same 93 descriptors and the same classifier moved from chance level to an AUC of 0.80 simply by changing the region from which they were extracted. This matters because the common design in this type of study, computing

everything from inside the expert-drawn annotation, is exactly the configuration that failed here. If our results generalize, they suggest that many existing radiomics pipelines may be looking in the wrong place. This result also indicates that patients are distinguished from controls by the relationship between the nerve and the tissue immediately surrounding it, rather than by the property of the nerve in isolation. That relationship is expressed at the boundary between the two, not within the nerve itself, so it is reasonable that features taken from just outside the annotation carried the signal, while features taken from inside it did not. The fixed-cube cropping strategy used in [12] is broadly consistent with this idea, since it also samples a region wider than the nerve itself; however, that study did not directly compare interior and surrounding regions, so this connection remains indirect.

The fixed-geometry control shows that annotation location alone carries some signal (AUC 0.633). Still, the gap between this control and the mask-derived shells (0.743–0.800) indicates that the shell result is not fully explained by location. Proximity to the actual annotation boundary appears to provide information that a region of fixed shape and size placed at the same center cannot capture. This control, therefore, rules out the possibility that the shell effect is purely a geometric artifact of how the annotations were drawn, while showing that location and boundary proximity both play a role. We would still recommend a fixed-geometry control as standard practice in similar studies, since it is the only way to disentangle these two contributions. This study has several limitations. The cohort is small (n = 80 after excluding CISS-acquired patients) and comes from a single center, as reflected in the fairly wide confidence intervals. A single reader produced all annotations, so we cannot yet say whether the shell effect would hold with annotations from a different reader. Because the fixed-geometry control operated below the mask-derived shells, we cannot fully separate content driven by proximity to the annotation boundary from the smaller residual contribution of the annotation shape itself; the control rules out a purely geometric artifact but not a partial one. No external, independent dataset was used to confirm the results. Finally, this study is an ablation experiment comparing regions using a fixed, deliberately simple pipeline, not an attempt to build an optimized diagnostic classifier; the absolute AUC values should therefore be read as a comparison between regions rather than as a claim about real-world diagnostic performance. Within these constraints, the aim of the proposed approach is not to serve as a primary diagnostic method in classical trigeminal neuralgia, which remains a clinical diagnosis. The principal role of imaging is to exclude secondary causes and to characterize neurovascular conflict before intervention. A quantitative peri-nerve marker, i.e., a shell-based feature set, may instead provide additional value for clinical management in the settings where visual assessment is uninformative, such as patients whose examinations are reported as unremarkable despite typical symptoms, and the well-recognized discrepancy in which neurovascular contact is present on the asymptomatic side or absent on the symptomatic one. A natural next step, which the present design cannot address, is a within-patient comparison of the symptomatic and asymptomatic sides, followed by testing whether peri-nerve features relate to response to microvascular decompression or radiosurgery. Until such data are available, the present findings should be interpreted as a methodological result concerning where to look, rather than as evidence that a peri-nerve signature is ready for clinical use.

VII. CONCLUSION

In this paper, we investigated where in a brain MRI the information needed for classification originates, using a small nerve structure as a case study. We compared classification performance across several spatial regions derived from the same expert annotation: the interior of the annotation itself, narrow bands immediately surrounding it, and a large box enclosing it. We also tested a fixed-geometry control region to determine whether the observed effect was due to image content or merely to the shape of the annotation. Intensity and texture predictors computed inside the expert-drawn annotation carried essentially no usable signal, performing no better than chance. The same predictors, computed instead in a narrow band just outside the annotation, reached an

AUC of 0.80. The fixed-geometry control region reached an AUC of 0.633, above chance, but noticeably lower than the shells derived from the annotation, showing that annotation location alone explains part, but not all, of this result; proximity to the true annotation boundary contributes additional signal beyond location alone. The large enclosing box, in contrast, performed poorly, showing that the benefit is specific to a fairly narrow region and is lost once the region becomes too large. These results indicate that, for this task, the region from which features are extracted is not a minor preprocessing choice but a design decision that can determine whether a model succeeds or fails. It deserves the same level of scrutiny and validation as the choice of features or classifier, rather than being fixed by convention to the expert annotation alone.

REFERENCES

[1]. G. Cruccu et al., "Trigeminal neuralgia: New classification and diagnostic grading for practice and research," Neurology, vol. 87, no. 2, pp. 220-228, 2016.
[2]. S. Maarbjerg, G. Di Stefano, L. Bendtsen, and G. Cruccu, "Trigeminal neuralgia-diagnosis and treatment," Cephalalgia, vol. 37, no. 7, pp. 648-657, 2017.
[3]. J. Cheng, J. Meng, W. Liu, H. Zhang, X. Hui, and D. Lei, "Nerve atrophy in trigeminal neuralgia due to neurovascular compression and its association with surgical outcomes after microvascular decompression," Acta Neurochir., vol. 159, no. 9, pp. 1699-1705, 2017.
[4]. D. P. Seeburg, B. Northcutt, N. Aygun, and A. M. Blitz, "The role of imaging for trigeminal neuralgia: A segmental approach to high-resolution MRI," Neurosurg. Clin. N. Am., vol. 27, no. 3, pp. 315-326, 2016.
[5]. R.H. Jani, M.A. Hughes, M.S. Gold, B.F. Branstetter, Z.E. Ligus, and R.F. Sekula Jr, "Trigeminal nerve compression without trigeminal neuralgia: Intraoperative vs imaging evidence," Neurosurg., vol. 84, no. 1, pp. 60-65, 2019.
[6]. R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," Radiology, vol. 278, no. 2, pp. 563-577, 2016.
[7]. P. Lambin et al., "Radiomics: The bridge between medical imaging and personalized medicine," Nat. Rev. Clin. Oncol., vol. 14, no. 12, pp. 749-762, 2017.
[8]. F. Wang, A. Ma, Z. Wu, M. Xie, P. Lun, and P. Sun, "Development and validation of radiomics models for the prediction of diagnosis of classic trigeminal neuralgia," Front. Neurosci., vol. 17, pp. 1188590, 2023.
[9]. K.L. Mulford, S.L. Moen, A.W. Grande, D.R. Nixdorf, and P.F. Van de Moortele, "Identifying symptomatic trigeminal nerves from MRI in a cohort of trigeminal neuralgia patients using radiomics," Neuroradiology, vol. 64, no. 3, pp.603-609, 2022.
[10]. H. Li et al., "A combined radiomics and anatomical features model enhances MRI-based recognition of symptomatic nerves in primary trigeminal neuralgia," Front—Neurosci., vol. 18, pp. 1500584, 2024.
[11]. J.H. Han et al., "Diagnosis of trigeminal neuralgia based on plain skull radiography using a convolutional neural network," Sci. Rep., vol. 15, no. 1, pp. 18888, 2025.
[12]. K.M. Halbert-Elliott, et al., "Deep learning-based segmentation of the trigeminal nerve and surrounding vasculature in trigeminal neuralgia," J. Neurosurg., vol. 143, no. 1, pp.83-91, 2025.
[13]. J.J. Van Griethuysen et al., "Computational radiomics system to decode the radiographic phenotype," Cancer Res., vol. 77, no. 21, pp. e104-e107, 2017.
[14]. A. Zwanenburg et al., "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," Radiology, vol. 295, no. 2, pp. 328-338, 2020.
[15]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn—Res., vol. 12, pp. 2825-2830, 2011.
[16]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861-874, 2006.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

From Classical Retrieval to LLM-Native and Multimodal Search: A Synthesis of an Ongoing Convergence

Fatou Kamara
Gaston Berger University
Laboratoire d'Analyse Numérique et Informatique
Saint-Louis, Senegal
fatou.kamara@ugb.edu.sn

Mouhamadou Lamine Diakhame
Gaston Berger University
Laboratoire d'Algèbre, de Cryptographie, de Codes et Applications
Saint-Louis, Senegal
diakhame.mouhamadou-lamine@ugb.edu.sn

Abstract
Information Retrieval (IR) has developed over more than seventy years, from lexical indexing and probabilistic weighting models to dense neural retrieval. Large Language Models (LLMs) are now reshaping the field in three ways: as components inside classical IR pipelines (query expansion, reranking); as end-to-end retrieval systems that internalize a corpus (generative and self-retrieval); and as the party that IR serves, through Retrieval-Augmented Generation (RAG), where retrieval becomes external memory for a model rather than a human searcher. A fourth dimension cuts across these: real-world content is increasingly multimodal text, images, audio, and video, which classical lexical IR was not built to index directly, but which the shared embedding space of dense retrieval extends to more naturally. This paper synthesizes recent work on the three integration modes and their multimodal extension. It identifies the open challenges common to all: the lack of standardized joint evaluation for RAG, the limited interpretability of neural matching, precision/recall trade-offs revived by hybrid dense-sparse methods, cross-modal alignment, and the computational and environmental cost of LLM-driven search. We give a structured reading of an active but fragmented area, organized around a single claim: LLMs do not so much replace IR as recompose four of its long-standing assumptions: what is indexed, what counts as a match, who is served, and which modality is retrieved.

Keywords: Information Retrieval; Large Language Models; Retrieval-Augmented Generation; Dense Retrieval; Generative Retrieval; Multimodal Retrieval; Neural Ranking; Evaluation

I. INTRODUCTION

Information Retrieval (IR) was formalized in the 1950s around a deliberately narrow problem: given a large collection of documents and a user query, return the documents most likely to meet the user's information need [1]. Early systems relied on term-overlap indexing [1]. The field then organized itself around three canonical retrieval models: Boolean, Vector Space, and Probabilistic [1],[2] validated through the Cranfield paradigm and, from 1992 onward, through NIST's Text Retrieval Conference (TREC) [1]. Decades of engineering on top of these models, most visibly the Divergence-From-Randomness (DFR) framework in the Terrier platform [3], have produced systems that are still used as reference baselines today. The 2010s brought a neural turn: queries and documents are embedded into continuous vector spaces, and toolkits such as Pyserini made hybrid sparse/dense pipelines reproducible at scale [4], [5].

LLMs are reshaping this trajectory again, but not along a single axis. Reviewing the recent literature, we see them entering IR in at least three structurally different ways:

- as *components* inserted into an otherwise classical pipeline for contextual query expansion, relevance-model extension, or feature-based reranking [6], [7], [8];

- as *end-to-end retrievers* that collapse indexing, matching, and ranking into a single generative model, echoing the broader shift toward generative retrieval [9, 10];
- as the *served party* in Retrieval-Augmented Generation (RAG), where IR becomes an external memory feeding an LLM rather than a human searcher [11-13].

A fourth, cross-cutting dimension compounds these three: real-world content is increasingly multimodal text, images, audio, video, and structured evidence such as tables or document screenshots. Classical lexical IR was not designed to index or match such content directly; it has handled it only indirectly, through manual annotation, captions, or other textual surrogates, rather than over the raw signal itself. The shared embedding-space paradigm of dense retrieval, by contrast, extends to it more naturally [14], [15].

Scope and method. This is a targeted, argument-driven synthesis, not a systematic review under a pre-registered protocol such as PRISMA: the aim is a defensible reading of a fast-moving area, favoring depth of evidence per claim over exhaustive coverage. A source was kept when it offered concrete, verifiable evidence, a quantitative result, an explicit taxonomy, or a foundational definition for at least one of the four dimensions, drawn mostly from peer-reviewed IR/NLP venues (ECIR, SIGIR, ICTIR, TREC, CORIA, IEEE Access), with arXiv preprints admitted only for the most recent work. The 19 sources fall into four strata: (1) foundational IR papers [1],[3]; (2) empirical papers instantiating each convergence mode [4], [6-9], [11]; (3) recent (20232025) surveys giving field-level taxonomies for generative retrieval, LLM-IR, and RAG evaluation [5], [10], [12], [13]; and (4) sources for the multimodal extension [14], [15], plus an expert-workshop report as an independent check [16]. The three-mode organization is not a new taxonomy; it reorganizes distinctions already present across the surveys in stratum (3) [10], [12], [13]; the contribution is the reading in Section 7, whose limitations are also stated in that section. This paper does not propose a new architecture; it offers a structured, literature-grounded reading of these four dimensions. Section 2 revisits the classical foundations later sections build on; Section 3 covers the dense/neural turn and the interpretability question it raised; Section 4 details the three integration modes as an organizing lens rather than a new taxonomy; Section 5 extends the analysis to multimodal retrieval; Section 6 draws together the cross-cutting open challenges; and Section 7 develops the paper's central claim that what is really changing is not a single technique but four of IR's core assumptions: what is indexed, what counts as a match, who the system serves, and what modality is retrieved.

II. CLASSICAL INFORMATION RETRIEVAL FOUNDATIONS

Classical IR rests on three canonical retrieval models that are still the vocabulary of the field: the *Boolean model*, the *Vector Space Model* (Salton), and the *Probabilistic model* of Robertson and Sparck Jones [1], [2]. All three share the same pipeline, summarized in Figure 1: documents are parsed into an inverted index, the query is processed through the same representation, a weighting function scores the document/query overlap, and a ranked list is returned, optionally refined by relevance feedback, which folds information from previously retrieved documents back into the query.

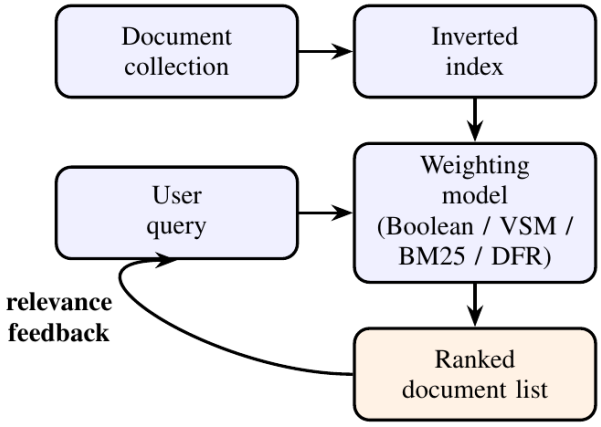


Figure 1. The classical IR pipeline behind the Boolean, Vector Space, and Probabilistic models: a collection is parsed into an inverted index; a query is scored by a weighting model (Boolean, VSM, BM25, or DFR); a ranked list is returned, optionally refined by relevance feedback. Later sections modify the weighting stage (Figure 2) or replace the pipeline (Figure 3). This architecture is not merely historical. The Terrier platform [3] turns the DFR framework, a generalization of Harter's 2-Poisson model, into more than fifty parameter-free document-weighting models alongside tf-idf, BM25, and language-modeling baselines, and it remains in active use (for example, as the retrieval backbone in later TREC Deep Learning submissions [8]). Its efficiency is notable: on the 426 GB TREC.GOV2: A Terabyte collection and a full Terrier index are only about 4.1% of the collection's size, and a long-query-with-expansion run reached a MAP of 0.3075 at the TREC 2004 Terabyte track, more than double the participants' median of 0.1427 [3]. Evaluation in this period is standardized around precision and recall, whose trade-off is already a defining tension of the field [1, 2]: precision-oriented systems were often justified by the observation that most users rarely look past the first results page [1]. That tension between exhaustive recall and user-facing precision returns, in a different guise, in hybrid dense-sparse retrieval (Section 6).

III. THE DENSE AND NEURAL RETRIEVAL TURN

Neural IR reorganizes the weighting stage of Figure 1 rather than the pipeline itself. Hambarde and Proença [5] group neural ranking models into three families: *representation-based* models that encode query and document independently before a shallow comparison (e.g., DSSM); *interaction-based* models that build fine-grained query-document interaction matrices before aggregating a score (e.g., ARC-I/II, MatchPyramid, K-NRM, DUET, PACRR, Conv-KNRM); and *hybrid* models. Pretrained transformers then displaced this generation of models: cross-encoder rankers such as monoBERT and duoBERT, and interaction-aware architectures such as CEDR, became standard second-stage rerankers [5].

Dense retrieval embeds queries and documents in a shared continuous vector space and scores candidates by vector similarity rather than lexical overlap. The Pyserini toolkit [4] made this reproducible at scale by wrapping Lucene-based sparse retrieval (via Anserini) and FAISS-based dense retrieval behind a single interface, along with a HybridSearcher that fuses sparse and dense scores (Figure 2). On the MS MARCO passage ranking dev set, Lin et al. report MRR@10 rising from 0.184 for default BM25 to 0.335 for a dense TCT-ColBERT retriever, 0.3530.365 for sparse-dense hybrids, and up to 0.420 for a multi-stage Expando-Mono-DuoT5 pipeline [4], evidence that dense and sparse signals are complementary rather than interchangeable, a point echoed in Section 6.

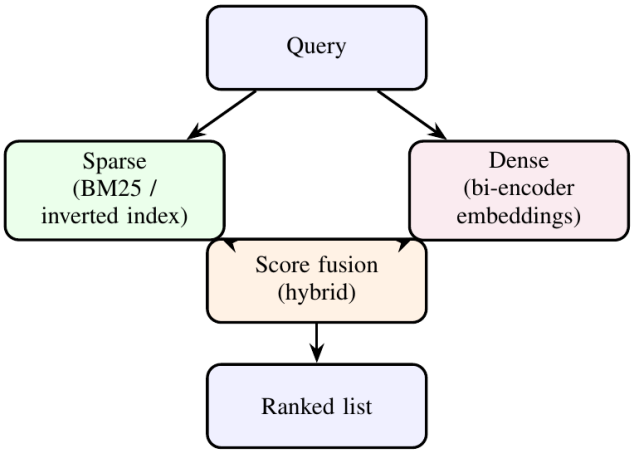


Figure 2. Sparse, dense, and hybrid first-stage retrieval: a query is scored by a sparse path (BM25 over an inverted index), and a dense path (bi-encoder embeddings), and the scores are fused into one ranked list, as in Pyserini's HybridSearcher [4]. On MS MARCO, the two signals prove complementary rather than substitutable (Section 3).

This gain in effectiveness came at the cost of transparency: interaction-based neural rankers are accurate, but their internal matching mechanism is largely opaque. Vast et al. study this directly, ablating attention pathways in monoBERT-style cross-encoders on the TREC Deep Learning 20192022 judgment sets [17]. Ablating the document's self-contextualization (documents that attend to themselves) yields the largest single-pathway drop in ranking quality. At the same time, the [CLS] token seems to aggregate relevance only in the later layers, playing a largely “no-op” role earlier. The same analysis finds that early-to-middle layers capture lexical, word-level correspondence between query and document, whereas later layers match semantically contextualized tokens [17]. This interpretability gap, effective matching without a clear account of why it matches, is one of the threads carried into the LLM era discussed next.

IV. THREE MODES OF LLMIR CONVERGENCE

Across the recent literature, LLMs do not enter IR through a single interface. We organize the convergence into three modes, shown together in Figure 3. These form a spectrum rather than a strict partition: a deployed RAG system (mode (c)) commonly embeds mode (a) components, such as LLM-based query rewriting and reranking, inside its retriever. The modes describe where the LLM does the decisive work; they are not mutually exclusive architectures.

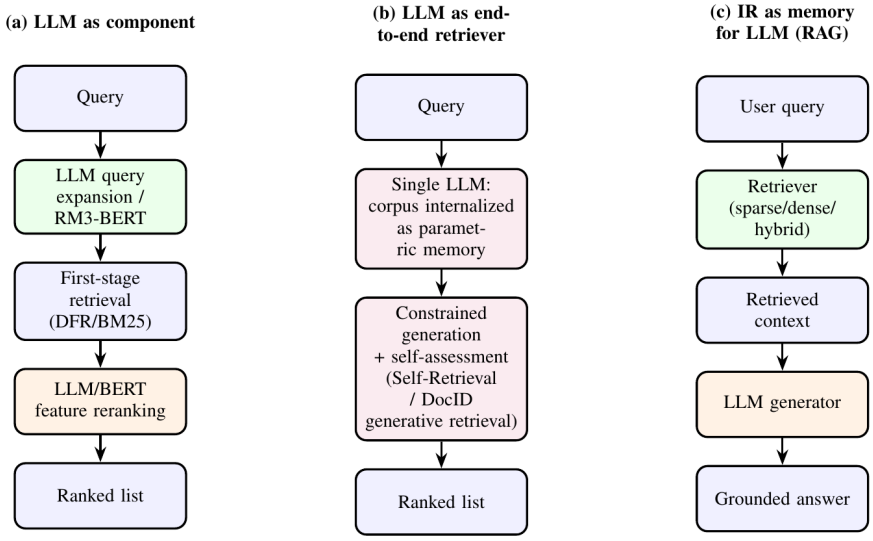


Figure 3. The three modes of LLMIR convergence: (a) LLMs augment a classical pipeline; (b) a single LLM performs indexing, retrieval, and ranking internally; (c) classical or dense IR becomes external memory serving an LLM's generation (RAG).

1. Mode (a): LLMs as Components Inside Classical Pipelines

The least disruptive integration keeps the pipeline in Figure 12 intact and inserts a language model at a single stage. Wang et al. extend pseudo-relevance feedback, classically formulated by Rocchio, to relevance models for multi-representation dense retrievers, refining the query's dense representation from an initial pseudo-relevant set rather than a bag-of-words query [6]. Kheloui et al. take a distinct route: their RM3-BERT model keeps the Relevance Model (RM3) as the backbone and linearly interpolates it with a BERT-based term-weighting component, tuned by a mixing parameter γ [7]. On AP88 and Robust-2004, RM3-BERT clearly outperforms RM3 alone, with MAP gains from 5.73% (short/title queries) up to 20.07% (long/description queries); the best mixture keeps γ=0.7, the lexical RM is weighted *more* than the BERT component, since pure BERT scoring (γ=0) does not beat RM3 on its own [7]. The catch is cost: BERT inference adds roughly 72 seconds per query on commodity hardware, which the authors flag as impractical for real-time search and a reason to move to pre-indexed, ColBERT-style representations. A further instance of mode (a) is feature-fusion reranking. The University of Glasgow Terrier team's TREC 2020 Deep Learning Track submission [8] keeps a three-stage architecture: DFR-based candidate retrieval (DPH) with Bo1 and T5-based query expansion; feature calculation (including BERT-[CLS], ColBERT, and SciBERT scores); and LambdaMART reranking. Notably, their single strongest feature across TREC 2019 and TREC 2020 was the ColBERT MaxPassage score, and their best-submitted run used no learning-to-rank reranking at all, due to the scarcity of training judgments relative to the complexity of the feature-based LTR model [8]. This is a useful counter-example to the assumption that adding more LLM-derived features always improves ranking: the contribution of each component still has to be validated empirically.

2. Mode (b): LLMs as End-to-End Retrievers

A more radical integration folds indexing, matching, and ranking into a single generative model. Generative retrieval recasts search as a sequence-to-sequence mapping from a query directly to a document identifier

(DocID), removing the separate inverted-index and reranking stages [10]. Kuo et al. organize the DocID design space into numerical identifiers (arbitrarily assigned or semantically structured) and string identifiers (substrings, titles/URLs, or pseudo-queries), and survey systems built around each of these, including DSI, DSI++, NCI, GENRE, SEAL, and CorpusBrain [10]. Self-Retrieval [9] goes further: one LLM internalizes the corpus in a self-supervised indexing phase, then retrieves by generate-then-self-assess, playing the role that indexing, matching, and reranking play separately in Figure 1. Mode (b) trades the transparency of an explicit index for the flexibility of a parametric one, at the cost of the interpretability and update guarantees inverted indexes give by construction, a trade-off this paper returns to in Section 6.

3. Mode (c): IR as External Memory for LLMs (RAG)

The third and most widely deployed mode reverses the direction of service. Instead of IR helping a human find documents, IR retrieves context that helps an LLM generate a better answer. Zhu et al. organize LLM-augmented IR around four modules: Query Rewriter, Retriever, Reranker, and Reader, plus an emerging Search-Agent layer that folds retrieval into the generation loop [12]. This raises a new evaluation problem: retrieval quality and generation quality are no longer separable. El-Yagouby et al. address it using a two-phase framework that combines a Longest-Common-Subsequence IR-quality score with an LLM-judge (GPT-4o) correctness rubric on HotpotQA, using a Llama3-8B generator across five retrievers [11]. A plain BM25 retriever gives the best IR score (0.62 at 1000 retrieved tokens), and while an LLM answering *without* any retrieval is fully incorrect 58% of the time, BM25-grounded RAG is fully correct 42% [11]. Two thresholds on the IR score (h=0.105, k=0.670) are then used to estimate answer quality without running the LLM, with roughly 56% of BM25 examples in the intermediate "hallucination risk" zone [11]. Gan et al.'s survey formalizes this joint problem, separating *internal* evaluation (retrieval and generation quality) from *external* evaluation (safety and efficiency), and catalogs metric families from MRR/NDCG/MAP and ROUGE/BLEU/BERTScore to LLM-as-judge protocols [13].

Table 1. The three modes of LLMIR convergence compared by strengths, limitations, and best-fit use cases: (a) an LLM inside a classical pipeline; (b) a single generative retriever; (c) IR as external memory for an LLM (RAG). Cited figures come from the studies in Section 4.

Mode	Strengths	Limitations/use cases
(a) Component	Keeps an auditable, updatable inverted index; incremental, measurable per-component gains (MAP +5.720% for RM3-BERT) [7]	Added latency (e.g., ~72 s/query for BERT-based RM3) [7]; gains not guaranteed best TREC 2020 run used no LTR at all [8]. Best suited to systems that already operate an inverted index.
(b) End-to-end	Unifies indexing, matching, and ranking in one model; flexible DocID design space [9], [10]	No explicit, inspectable index; unclear incremental-update story when new documents arrive [10]. Best suited to closed or slowly-changing corpora, and to research prototypes.
(c) RAG	Decouples parametric knowledge from external, up-to-date facts; even a plain BM25 retriever sharply reduces fully-incorrect answers, raising the fully-correct rate to 42%	Retrieval and generation quality are not yet jointly and consistently measured [13]; ~56% of BM25-based answers still fall within an ambiguous risk zone [11]. Best suited to open-domain QA and knowledge

Mode	Strengths	Limitations/use cases
	where the same LLM with no retrieval is fully incorrect 58% of the time [11]	bases that change faster than any model can be retrained.

V. THE MULTIMODAL EXTENSION

The three modes were formulated for text, but real content is rarely text-only, and the same structure reappears across modalities. Mei et al. frame this as an evolution from *MRAG1.0*, a "pseudo-multimodal" setting where non-textual content is first converted to text and then run through a standard text pipeline, to MRAG2.0, where the original content is kept, and a multimodal LLM (MLLM) interprets and generates directly on it [14]. Abootorabi et al. give a finer-grained taxonomy organized around retrieval strategies, fusion mechanisms, augmentation and generation techniques, and training strategies [15]. Concrete systems fill each branch: ColPali retrieves directly over page images, bypassing OCR; MuRAG and RA-CM3 jointly retrieve and generate across text and images; ColBERT/PreFLMR extend late interaction to multimodal retrieval; and CRAG adds retrieval-quality-aware correction inside the RAG loop [15].

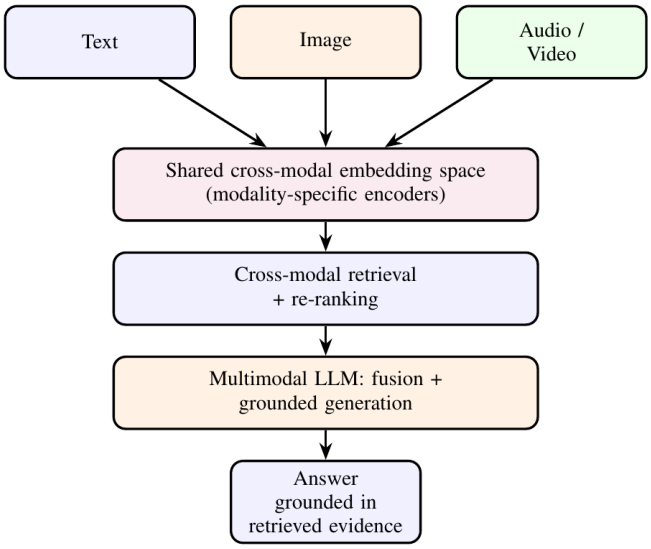


Figure 4. Multimodal RAG (MRAG2.0-style): modality-specific encoders feed a shared embedding space, from which cross-modal evidence is retrieved and grounded by a multimodal LLM [14], [15].

This extension (Figure 4) inherits the same three-mode structure as Section 4 multimodal encoders as components, end-to-end multimodal retrievers, and multimodal RAG. Still, it introduces cross-modal difficulties: aligning heterogeneous embedding geometries, a relative scarcity of multimodal benchmarks compared with decades of text-only TREC-style collections, and the higher cost of multimodal encoders [14], [15]. Badamasi et al. independently flag heterogeneous, multimodal data as a top-level challenge for information access even outside the RAG literature [2], a sign that this is a broadly recognized difficulty rather than one community's framing.

VI. CROSS-CUTTING OPEN CHALLENGES

Table 2 lists five challenges that recur, in different guises, across the modes and the multimodal extension.

Table 2. Five open challenges recurring across the three modes and the multimodal extension, each with the setting in which it surfaces. Directions and mitigations follow in the text.

Challenge	Where it surfaces
Evaluation standardization	Retrieval and generation quality must be assessed jointly in RAG [11]; no consensus metric set yet, though internal/external taxonomies are emerging [13].
Interpretability of matching	Neural cross-encoders match effectively but opaquely; specific attention heads carry lexical vs. semantic signal, with document self-contextualization the most critical single pathway [17].
Precision/recall trade-offs	Hybrid dense-sparse fusion can raise Recall@10 by up to 580% relative (13.9%→80.8% on MS MARCO) but at a cost in ranking quality (NDCG@10 0.83→0.51), and the gain vanishes on specialized-vocabulary corpora [18].
Cross-modal alignment	Modality-specific embedding geometries and scarce multimodal benchmarks limit generalization of dense-retrieval fusion across text/image/audio/video [14], [15].
Computational/environmental cost	BERT-scale inference can add tens of seconds per query on commodity hardware [7]; data-center energy and cooling for large-scale AI search raise resource-consumption concerns, which have been documented primarily in practitioner sources [19] and are still awaiting authoritative, peer-reviewed empirical quantification for LLM-driven retrieval.

The precision/recall tension is worth expanding on, since it revisits the classical trade-off in Section 2. That sparse and dense signals are complementary is established in peer-reviewed work (Section 3) [4]; the magnitudes below come from a non-peer-reviewed preprint [18] and are best read as illustrative. Weinberg's hybrid method, combining a compact bi-encoder with BM25 through weighted score fusion, reports Recall@10 rising from 13.9% (dense-only) to 80.8% (hybrid) on MS MARCO, ahead of a ColBERT-style baseline (15.0%) and a SPLADE-inspired learned-sparse baseline (4.2%) [18]. Yet NDCG@10 falls from 0.83 to 0.51, the recall/precision-quality tension resurfacing inside one fused score, and the advantage disappears on BEIR NFCorpus, where dense-only and hybrid recall are indistinguishable (17.6% vs. 17.4%), attributed to that corpus's specialized vocabulary and small size [18]. The gain is real but corpus-dependent, and bought at a measurable cost elsewhere in the ranking.

At a community level, the 2024 Computing Community Consortium workshop report [16], convening 44 IR, NLP, HCI, and AI researchers, groups future priorities into eight themes that closely mirror the challenges above: evaluation; training and reasoning; modeling users; social ramifications; personalization; efficiency and scalability; AI agents; and foundation models for information access [16]. That an independent expert workshop converges with the challenges surfaced bottom-up here supports treating them as structural open problems rather than incidental limitations of specific systems.

Directions and mitigations. The challenges of Table 2 and the per-mode limitations of Table 1 are not equally intractable. For *evaluation*, the most actionable near-term direction is automated, reference-free assessment: El-Yagouby et al.'s two-threshold scheme predicts RAG answer quality from a retrieval-quality score [11], and Gan et al.'s internal/external decomposition offers a common reporting vocabulary [13]. Both rely on *LLM-as-a-*

Judge protocols, whose known failure modes, position and verbosity biases, prompt sensitivity, and drift when the judge is updated mean a judge should be calibrated against human labels and reported with its model and version fixed [13]. For the *mode-specific* limitations, three directions recur: a high-recall-then-rerank pipeline mitigates both the ranking-quality cost of hybrid fusion and the RAG "risk zone" [18]; pre-computed ColBERT-style representations cut the per-query latency of BERT-based components in mode (a) [7], [8]; and incremental indexing remains the prerequisite for mode (b) to serve fast-changing corpora [10].

VII. DISCUSSION: RECOMPOSING IR'S CORE ASSUMPTIONS

Reading Sections 25 together, the changes LLMs introduce are not well described as "adding a language model to IR." They are better described as a recomposition of four assumptions that classical IR treated as fixed, summarized in Table 3.

Table 3. The paper's central claim: four assumptions classical IR treated as fixed (middle) and what each becomes under LLM-native and multimodal retrieval (right), what is indexed, what counts as a match, who is served, and what modality is retrieved.

Axis	Classical IR	LLM-native / multimodal IR
What is indexed	Terms in an inverted index [3]	DocIDs in a parametric model [9], [10] or cross-modal embeddings [14]
What counts as a match	Lexical / statistical overlap (BM25, DFR) [3]	Learned semantic similarity, often uninterpretable [5], [17]
Who is served	A human searcher [1]	An LLM consumer, via RAG [11], [12]
What modality	Text [1]	Text, image, audio, video, jointly [15]

Each recomposition carries its own unresolved question. Moving what is indexed from an inverted structure to a model's parameters (mode (b)) trades updateability and inspectability for flexibility. A DocID-generating LLM cannot be updated by appending to an index the way Terrier's structure can [3, 10]. Moving what counts as a match from lexical overlap to learned similarity buys effectiveness (Section 3) at the cost of interpretability, which is only partially recoverable by post hoc analysis [17]. Moving who is served from a human to an LLM (mode (c)) is the most consequential shift, since it changes what "relevance" is optimized for: in the HotpotQA setup above, El-Yagouby et al. find plain BM25 the strongest of five retrievers for downstream answer correctness [11] a single-benchmark result that should not be over-read, but a warning that retrieval metrics tuned for human satisfaction may not carry over to LLM consumption. Finally, moving from a text-only modality to a joint text/image/audio/video space is the least mature axis, constrained more by scarcity of benchmarks and alignments than by any architectural obstacle [14], [15].

A dimension that cuts across all four axes, only partly addressed by the sources here, is language. *What counts as a match* changes most: cross-lingual dense retrieval can align query and document across languages in one embedding space, impossible for lexical overlap, but degrades where training data is thinnest. *What is indexed* depends on a tokenizer and parametric memory covering the target script, which works much better for high-resource languages. *Who is served* inherits the retriever's coverage, so the hallucination-reduction grounding in English need not transfer to low-resource languages. And *what modality is retrieved* compounds the scarcity of benchmark data, since jointly multilingual and multimodal collections are rarer still [15]; heterogeneous,

unevenly resourced content is itself flagged as a first-order challenge [2]. Cross-lingual and low-resource robustness is thus a stress test that the four recomposed assumptions have yet to pass uniformly.

1. Limitations of This Survey

Three limitations bear on how these claims should be read. First, source selection was targeted rather than systematic (Section 4), so this is a reading of the space, not a census. Second, a few sources sit outside mainstream peer-reviewed IR/NLP publishing the hybrid dense-sparse report [18], an industry preprint posted to SSRN, and a practitioner survey of AI-based search [19] kept for their concrete quantitative case studies but carrying less evidentiary weight than the TREC-, SIGIR-, or ECIR-anchored studies, and flagged as such throughout. Third, coverage is skewed over time: most LLM-IR and RAG sources are from 20232025, and pre-2020 neural IR is represented mainly by a 2023 survey [5] rather than by the primary sources it summarizes.

VIII. CONCLUSION

Seventy years after it was formalized, Information Retrieval is not being replaced by Large Language Models so much as recomposed by them, along at least four axes: what is indexed, what counts as a match, whom the system serves, and what modality is retrieved. We organized the recent literature around three integration modes: LLMs as pipeline components, LLMs as end-to-end retrievers, and IR as external memory for LLMs via RAG, and showed that each extends naturally but not painlessly to multimodal content. The open challenges standardized joint evaluation of retrieval and generation, interpretable neural matching, corpus-dependent precision/recall trade-offs, cross-modal alignment, and the computational and environmental cost of LLM-driven search are corroborated both by individual studies and by a community-wide expert workshop [16], which suggests they are structural rather than incidental. Clarifying where classical IR foundations still hold, and where they do not, is a necessary step before the field can agree on shared evaluation practices for this new generation of retrieval systems.

REFERENCES

[1]. S. O. Kuyoro, A. Oludele, F. A. Ibikunle, and S. B. Abel, "Information Retrieval: An Overview," International Journal of Advanced Research in Computer Science, vol. 3, no. 5, pp. 175178, 2012.

[2]. B. Badamasi, Y. Galadima, J. M. Abdulkarim, J. U. Aliyu, R. Umar, and F. Hosea, "Qualitative Study on Advancing Information Access: The Role of Information Retrieval Systems in Enhancing Digital User Experience," Zamfara International Journal of Education, vol. 5, no. 3, pp. 190195, 2025, doi: 10.64348/zije. 202555.

[3]. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson, "Terrier Information Retrieval Platform," in Proc. 27th European Conf. on Information Retrieval (ECIR), LNCS vol. 3408, 2005, pp. 517519.

[4]. J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira, "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations," in Proc. 44th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2021, pp. 23562362.

[5]. K. A. Hambarde and H. Proença, "Information Retrieval: Recent Advances and Beyond," IEEE Access, 2023, doi: 10.1109/ACCESS.2023.3295776.

[6]. X. Wang, C. Macdonald, N. Tonellotto, and I. Ounis, "Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval," in Proc. ACM SIGIR Int. Conf. on Theory of Information Retrieval (ICTIR), 2021, pp. 297306.

[7]. A. Kheloui, A. Hammache et M. Boughanem, "Extension du modèle de pertinence avec BERT", dans Actes de la 19e Conférence en Recherche d'Information et Applications (CORIA), 2024, doi: 10.24348/coria.2024.long_20.
[8]. X. Wang, Y. Wu, X. Wang, C. Macdonald, and I. Ounis, "University of Glasgow Terrier Team at the TREC 2020 Deep Learning Track," in Proc. 29th Text REtrieval Conf. (TREC), 2020.
[9]. Q. Tang, J. Chen, B. Yu, Y. Lu, C. Fu, H. Yu, H. Lin, F. Huang, B. He, X. Han, L. Sun, and Y. Li, "Self-Retrieval: Building an Information Retrieval System with One Large Language Model," arXiv:2403.00801, 2024.
[10]. T.-L. Kuo, T.-W. Chiu, T.-S. Lin, S.-Y. Wu, C.-W. Huang, and Y.-N. Chen, "A Survey of Generative Information Retrieval," arXiv:2406.01197, 2024.
[11]. M.-A. El-Yagouby, P. Mulhem, J.-P. Chevallet, and É. Gaussier, "Cadre d'évaluation pour les systèmes de génération augmentée (RAG): combinaison des performances de recherche d'information et de LLM," dans Actes de la 20e Conférence en Recherche d'Information et Applications (CORIA), Marseille, France, 2025, pp. 5060.
[12]. Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, and J.-R. Wen, "Large Language Models for Information Retrieval: A Survey," arXiv:2308.07107, 2023.
[13]. A. Gan, H. Yu, K. Zhang, Q. Liu, W. Yan, Z. Huang, S. Tong, and G. Hu, "Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey," arXiv:2504.14891, 2025.
[14]. L. Mei, S. Mo, Z. Yang, and C. Chen, "A Survey of Multimodal Retrieval-Augmented Generation," arXiv:2504.08748, 2025.
[15]. M. M. Abootorabi, A. Zobeiri, M. Dehghani, M. Mohammadkhani, B. Mohammadi, O. Ghahroodi, M. S. Baghshah, and E. Asgari, "Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation," arXiv:2502.08826, 2025.
[16]. J. Allan, E. Choi, D. P. Lopresti, and H. Zamani, "Future of Information Retrieval Research in the Age of Generative AI," Computing Community Consortium (CCC) Workshop Report, Computing Research Association, arXiv:2412.02043, Dec. 2024.
[17]. M. Vast, B. Van Cooten, L. Soulier, and B. Piwowarski, "Understanding Matching Mechanisms in Cross-Encoders," in Proc. Workshop on Explainability in Information Retrieval (WExIR) at SIGIR 2025, Padua, Italy, 2025, arXiv:2507.14604.
[18]. A. I. Weinberg, "Hybrid Dense-Sparse Retrieval for High-Recall Information Retrieval," SSRN preprint 6013934, 2026.
[19]. O. Segeda, "Building Intelligent Search Systems: Advances in AI-Based Information Retrieval," The American Journal of Applied Sciences, vol. 7, no. 6, pp. 611, 2025, doi: 10.37547/tajas/Volume07Issue06-02.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Bibliometric Analysis of Artificial Intelligence Assisted Medical Imaging Research: Global Trends and Turkey's Position (2021-2025)

Kemal AKKANAT
Dept. of Computer Engineering
Trakya University
Edirne, Türkiye
kemalakkanat@trakya.edu.tr
0000-0002-6448-7378

Muhittin TANOBA
Dept. of Computer Engineering
Trakya University
Edirne, Türkiye
muhittintanoba@trakya.edu.tr
0009-0002-6641-759X

Abstract
Between 2021 and 2025, artificial intelligence (AI) and deep learning (DL) methods came to play a defining role in medical imaging research. This study sets a global publication corpus drawn from Scopus (19,968 records after deduplication) against a Turkey-affiliated subset (4,843 records), mapping research trends, thematic shifts, and the keyword clusters that dominate each at both the thematic-group and individual-keyword levels. The bibliometrix R package and VOSviewer were used to build keyword co-occurrence networks and Sankey-based thematic evolution diagrams. CNN and deep learning still anchor the global literature. That dominance, however, is no longer unchallenged: between 2024 and 2025, vision transformers (ViT), explainable AI (XAI) methods such as Grad-CAM, and the broader push for interpretability reshaped the field's center of gravity. The Turkish dataset tells a different story: machine learning and classification lead the field and, once frequencies are normalized by publication volume, interest in ChatGPT and related large-language-model terms, once grouped into a single category, sits at a markedly higher relative density (26.2×, χ^2=290.2, $p<0.001$) than in the global corpus. A supplementary, ungrouped keyword analysis tentatively identifies clinical decision support as an early point of convergence between the two literatures in 2024-2025.

Keywords: bibliometric analysis, deep learning, medical imaging, VOSviewer, bibliometrix, Turkey, thematic evolution

I. INTRODUCTION

Deep learning's move into medical imaging has been one of the fastest technological shifts of the past five years. Convolutional neural networks (CNNs) outpaced traditional pattern-recognition methods. They gave diagnostic support systems a solid technical foundation [1], [2], with the same architecture continuing to prove effective on general image-classification tasks outside the medical domain, including semantic-communication-based classification pipelines [3], [4]. Residual connections [5], dense connectivity [6], and EfficientNet's compound scaling [7] later enabled CNNs to go deeper and more cheaply without sacrificing trainability. The Vision Transformer (ViT) architecture, proposed in 2021, drew researchers' attention to attention-based approaches that split images into patches and model global relationships among them, rather than relying on CNNs' local, hierarchical receptive fields [8]. Bibliometric analysis offers a way to track which methods are actually setting the agenda. By processing large volumes of publication data with quantitative methods, researchers can map the structure of a field [9]. Donthu et al. [10] note that the method does double duty, measuring performance while also mapping the science itself, and that it travels well across disciplines. Chen et al. [11] examined 562 highly cited WoS articles from 2013–2023 and found CNN, classification, and segmentation dominant, but their window closes before the post-2024 rise of ViT and XAI. Wang et al. [12] reported that breast, lung, and skin cancer dominate cancer-focused DL applications, while Lin et al. [13] identified country-level productivity leaders between 2019 and 2023; neither examined Turkey specifically. This paper fills that gap with four

contributions: a comparative global-versus-Turkey analysis; a mapping of the research agenda through keyword networks, Sankey diagrams; a volume-normalized assessment of where Turkey stands relative to the global agenda, not merely a raw-count one; and a side-by-side look at ViT, XAI/Grad-CAM, and ChatGPT-related terms, supplemented by an ungrouped keyword-level pass that surfaces micro-themes too small to register thematically. The remainder of this paper is organized as follows: Section II reviews related bibliometric studies; Section III describes the corpus construction and analytical methods; Section IV presents the results, including keyword frequencies, co-occurrence networks, and thematic evolution; Section V discusses the findings; and Section VI concludes with limitations and future work.

II. RELATED WORK

Bibliometric studies at the intersection of medical imaging and AI have multiplied quickly (Table 1). Chen et al. [11] used CiteSpace to scan 2013–2023 and confirmed CNN-centered themes, but missed the newer ViT and XAI threads. Wang et al. [12] and Lin et al. [13] contributed findings on cancer applications and country-level productivity, respectively, but neither examined Turkey on its own terms. Litjens et al. [14], Shen et al. [15], and Miotto et al. [16] published broad reviews spanning pathology, radiology, and clinical decision support, while Dosovitskiy et al. [8] showed that ViT could rival or exceed CNNs with large-scale pretraining. Selvaraju et al.'s Grad-CAM method [17] helps explain why interpretability research has accelerated so fast, addressing a transparency gap in the field. Broader application reviews, such as Houssein et al.'s survey of deep and machine learning techniques for breast cancer imaging [18], implicitly highlight this same gap by cataloging the many black-box models now in clinical use. None of the existing literature provides a comparative, volume-normalized look at global and Turkey-affiliated publications across 2021–2025; this is the gap this study addresses.

Table 1: Comparison of Related Studies

Study	Scope	Period	Turkey-specific
Chen et al. [11]	CNN, classification, segmentation	2013–2023	No
Wang et al. [12]	Cancer-focused DL applications	2015–2023	No
Lin et al. [13]	Country-level AI productivity	2019–2023	No
This study	Global vs. Turkey, volume-normalized, ViT/XAI/LLM emergence	2021–2025	Yes

III. MATERIALS AND METHODS

1. Corpus Construction and Sampling

Two publications set the anchor for this study. The global dataset was derived from a Boolean query run in Scopus's TITLE-ABS-KEY field, combining AI-architecture terms with health, medical, clinical, radiology,

pathology, and imaging terms, and restricted to the computer science and medicine subject areas from 2021 to 2025. The exact query used was: TITLE-ABS-KEY(("deep learning" OR "machine learning" OR "convolutional neural network" OR "CNN" OR "transformer" OR "vision transformer" OR "ViT" OR "U-Net" OR "ResNet" OR "LSTM" OR "RNN" OR "GAN" OR "generative adversarial" OR "attention mechanism" OR "transfer learning") AND ("healthcare" OR "medical" OR "clinical" OR "hospital" OR "patient" OR "disease" OR "diagnosis" OR "radiology" OR "pathology" OR "imaging")) AND NOT TITLE-ABS-KEY("plant disease*" OR "plant detect*" OR "crop*" OR "agriculture" OR "soil*" OR "weed*" OR "pest*" OR "livestock") AND PUBYEAR > 2020 AND PUBYEAR < 2026 AND (LIMIT-TO(SUBJAREA, "COMP") OR LIMIT-TO(SUBJAREA, "MEDI")) AND (LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "re") OR LIMIT-TO(DOCTYPE, "cp")). The Turkey subset applied the same query with an additional AND (LIMIT-TO(AFFILCOUNTRY, "Turkey")) clause, capturing 4,847 records, 4,843 after deduplication. The global query returned 278,252 records in total, rising steadily from 32,639 in 2021 to 84,720 in 2025. Scopus caps single-query exports at 20,000 records. For the global corpus, the top 4,000 relevance-ranked records per year were therefore retained, yielding a 20,000-record working corpus. After DOI and title deduplication, the working corpora settled at 19,968 global records and 4,843 Turkey records. Processing the full frame was not feasible: Scopus's export interface is capped at 20,000 records in total and offers no random sampling option, and Bibliometrix's own memory ceiling is roughly the same. Together, these fixed the working corpus at 20,000 records, hence the 4,000-record-per-year allocation across the five years (a 50,000-record ceiling would instead have allowed 10,000 per year). This is a relevance-ranked, year-stratified truncation, not systematic sampling: within each year, Scopus's proprietary relevance score, rather than a random interval, determines inclusion, which means highly cited, metadata-complete records are overrepresented relative to low-citation, very recent, or non-English/national-journal records. Although the truncation procedure is non-probabilistic, its empirical regularity was assessed across five sequential 4,000-record slices, using Spearman's ρ and Jaccard's J over the top-20 keyword set (Table 9). Consecutive-slice stability was high (mean $\rho = 0.890$; mean $J = 0.741$), indicating strong thematic consistency despite the non-probabilistic truncation. Cumulative comparison against the first slice showed expected attenuation ($\rho = 0.488$; $J = 0.538$), consistent with gradual thematic drift. A chi-square test was significant at every step, but is known to be oversensitive at this sample size, so Spearman/Jaccard were used as the primary criteria. A Mann-Whitney U test found no difference across the first three transitions but a significant one between the fourth and fifth slices ($p = 0.0028$), consistent with the fifth slice (2025) having had less time to accrue citations. Taken together, these findings suggest the truncation, though not systematic by design, produced a stable and representative working corpus.

Records were included if they were published between 2021 and 2025, indexed in Scopus, written in English, and classified as an original article, review, or conference paper, and referencing the intersection of AI and medical imaging in the title, abstract, or keywords. Duplicates, retractions, non-clinical engineering studies, and records lacking an accessible abstract were excluded. Two independent researchers applied these criteria and resolved disagreements by discussion.

2. Analytical Tools and Parameters

Analysis ran on two platforms. Aria and Cuccurullo's [9] bibliometrix package, version 4.3.x, handled publication-trend analysis, country and institution classification, and keyword frequency counts at two levels, 23 canonical thematic groups such as Deep Learning, CNN, ViT, XAI, and ChatGPT/LLM, plus a supplementary pass across individual, ungrouped Author Keywords meant to catch early-stage micro-themes falling below any single group's detection threshold. Keyword standardization followed a controlled synonym-mapping protocol that unified plural, hyphenated, and near-duplicate variants into a single canonical form, with 10% of mappings cross-checked by a second researcher. Van Eck and Waltman's [19] VOSviewer, version

1.6.20, built keyword co-occurrence networks, clustered them with the Leiden algorithm at a resolution of 1.0, and generated overlay visualizations keyed to the average publication year. The co-occurrence parameters were derived from pilot runs rather than the default settings. These runs settled on a minimum frequency of 10 for the global set, and 5 for the Turkey set, a cap of 150 terms, and association strength as the normalization method. Thematic evolution used Bibliometrix's thematicEvolution(). Supplementary keyword terms needed at least 3 occurrences in the 2024 baseline year to enter the share-growth ranking. Because the global corpus is roughly 4.1 times Turkey's, every comparison is reported as both a raw count and a rate per 1,000 publications. All figures were produced in R 4.4.0. To assess whether Turkey/Global density differences were statistically significant, chi-square tests of independence were additionally computed for each keyword in Table 2, comparing Turkey presence/absence against the Global corpus (α=0.05).

IV. RESULTS

Table 2 reports raw frequencies alongside volume-normalized densities, with the Turkey/Global ratio in the final column (1.0 = equal relative weight; below 1 = Turkey under-use; above 1 = over-use). Two entries dominate that column: a grouped ChatGPT/LLM category at 26.2× (χ^2=290.2, p<0.001) and radiomics at 2.42× (χ^2=40.1, p<0.001), both indicating that Turkey is proportionally more engaged than the global literature. All differences in Table 2 are statistically significant (p<0.05) except Explainable AI (χ^2=3.3, p=0.070), where Turkey's apparent under-use is not distinguishable from the global rate. At the other extreme, the convolutional neural network sits at 0.23×, and the vision transformer at 0.22×. Turkey engages with these foundational architectures at under a quarter of the global rate.

Table 2: Global vs. Turkey Keyword Frequencies, Raw and Volume-Normalized

Keyword	Global f	Global /1000 pub	Turkey f	Turkey /1000 pub	Turkey/ Global ratio	χ^2	p
Deep learning	9,579	479.7	1,422	293.6	0.61x	546.8	<0.001
Convolutional neural network	3,138	157.2	177	36.5	0.23x	488.9	<0.001
Machine learning	2,688	134.6	1,311	270.6	2.01x	532.5	<0.001
Transfer learning	2,286	114.5	258	53.3	0.47x	158.1	<0.001
U-Net	627	31.4	65	13.4	0.43x	45.8	<0.001
MRI	574	28.7	80	16.5	0.57x	22.3	<0.001
Vision transformer	970	48.6	51	10.5	0.22x	142.1	<0.001
Explainable AI	184	9.2	31	6.4	0.69x	3.3	0.070
Grad-CAM	144	7.2	20	4.1	0.57x	5.2	0.023
Artificial intelligence	1,229	61.6	614	126.8	2.06x	240.1	<0.001
Radiomics	133	6.7	78	16.1	2.42x	40.1	<0.001
ChatGPT	3	0.2	57	11.8	78.32x	213.3	<0.001
LLM-grouped	14	0.7	89	18.4	26.20x	290.2	<0.001

1. Co-occurrence Networks

Figure 1, the global co-occurrence network, and Figure 2, the Turkey co-occurrence network, show the cluster-colored VOSviewer maps. Globally, “deep learning” sits at the center, branching to “CNN,” “breast cancer,” “COVID-19,” and “U-Net,” whose segmentation role traces to Ronneberger et al.’s original architecture [20], with the network splitting into an imaging/clinical cluster and a bioelectric-signal cluster. Turkey’s network organizes around four hubs: “deep learning,” “classification,” “transfer learning,” and “artificial intelligence,” with the “U-Net”/“MRI” clusters that loom large globally, barely visible here, matching their sub-1.0 density ratios in Table 2.

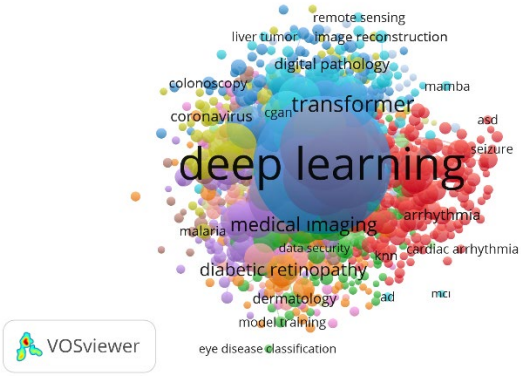


Figure 1. Global co-occurrence network

Figure 2 presents the corresponding network built from the Turkey-affiliated subset.

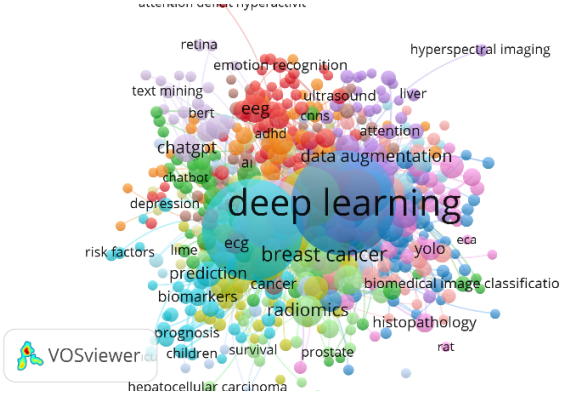


Figure 2. Turkey co-occurrence network

To examine whether architectural choices vary by imaging modality, Tables 3 and 4 cross-tabulate the five leading architectures (CNN, ViT, U-Net, ResNet, GAN) with seven imaging modalities. Globally, CNN-MRI is the dominant pairing (f=482), consistent with CNN's foundational role in neuroimaging. In the Turkish corpus, this pairing drops sharply (f=20). At the same time, ViT-histopathology (f=6) approaches ViT-MRI (f=3) in frequency — a modality distribution not visible in the aggregate architecture and modality counts reported separately in Table 2.

Table 3: Architecture × Modality Crosstab — Global (publication count)

Architecture	MRI	CT	X-ray	Ultrasound	Histopath.	Endoscopy	Dermoscopy
CNN	482	135	264	44	65	22	32
ViT	65	14	48	12	29	5	8
U-Net	58	17	10	6	7	1	1
ResNet	14	6	15	3	5	0	2
GAN	75	34	21	7	9	3	2

Table 4 repeats this breakdown for the Turkey-affiliated subset.

Table 4: Architecture × Modality Crosstab — Turkey (publication count)

Architecture	MRI	CT	X-ray	Ultrasound	Histopath.	Endoscopy	Dermoscopy
CNN	20	7	9	2	9	2	6
ViT	3	0	4	0	6	1	0
U-Net	10	1	0	1	2	0	1
ResNet	0	1	1	0	1	0	0
GAN	3	2	0	1	0	0	0

2. Overlay Visualizations

In Figure 3, the global overlay, and in Figure 4, the Turkey overlay, each node is colored by average publication year. Globally, “vision transformer,” “swin transformer,” and “radiomics” sit in the 2023–2024 band, confirming their recent surge, while older terms like “CNN” and “VGG16” stay in earlier tones. Turkey shows “radiomics,” “ChatGPT,” and “explainable AI” in similarly recent shades, but their clusters are visibly smaller, gaining currency, not yet critical mass. Because the 2024–2025 nodes in both figures overlap too densely to read individually, Tables 5 and 6 list the highest-recency keywords by average publication year for each network, along with their frequencies.

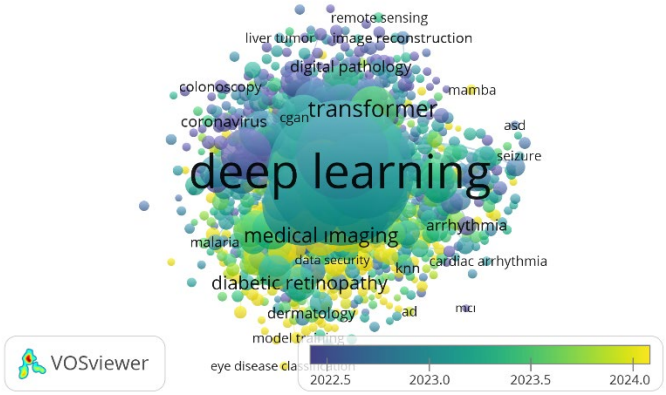


Figure 3. Global overlay network (by average publication year)

Figure 4 shows the same overlay computed for the Turkey-affiliated subset.

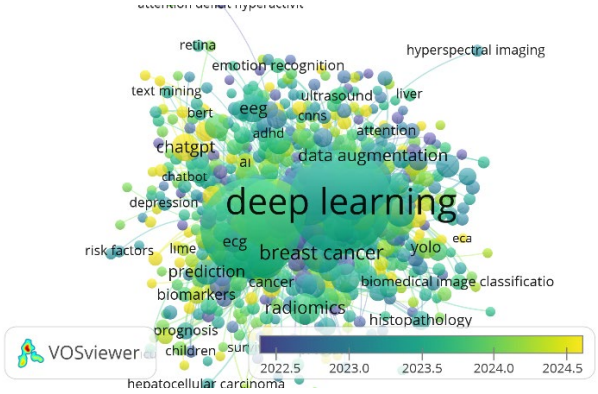


Figure 4. Turkey overlay network (by average publication year)

These frequency and average-publication-year values are reported in Tables 5 and 6 below.

Table 5: Top Recently-Emerging Keywords by Average Publication Year — Global Overlay Network (Figure 3)

Keyword	f	Avg. publication year
ChatGPT/LLM (grouped)	14	2024.2
Vision transformer	1425	2024.1
Swin transformer	126	2024.1
Grad-CAM	159	2024.1
Explainable AI	281	2024.0
Medical imaging	1532	2023.8
Anomaly detection	48	2023.6
VGG16	267	2023.5
Transformer	834	2023.5
ResNet	401	2023.4

Table 6 presents the equivalent ranking for the Turkey overlay network.

Table 6: Top Recently-Emerging Keywords by Average Publication Year — Turkey Overlay Network (Figure 4)

Keyword	f	Avg. publication year
ChatGPT/LLM (grouped)	89	2024.5
Vision transformer	85	2024.5
Swin transformer	15	2024.3
Explainable AI	92	2024.3

Keyword	f	Avg. publication year
Anomaly detection	17	2024.2
Grad-CAM	21	2024.1
Medical imaging	116	2024.1
Transformer	64	2023.9
Artificial intelligence	627	2023.9
Radiomics	78	2023.7

3. Thematic Evolution

Figure 5 shows the global thematic evolution as a Sankey diagram, and Figure 6 shows the same view for Turkey, tracing how themes shifted year by year. Globally, 2021 was dominated by "skin cancer," anchored early by Esteva et al.'s dermatologist-level classifier [21], alongside "deep learning" and "CNN"; "COVID-19" gained visibility in 2022; "transfer learning" and "transformer" emerged in 2023; and 2024 brought "breast cancer" and "vision transformer" to the fore. In Turkey, "deep learning" led every year from 2021 through 2024. Still, in 2025, "machine learning" overtook it for the first time (488 vs. 482 occurrences), a reversal not previously visible under the coarser multi-year binning used in Figure 6. "Classification," "explainable AI," and "anomaly detection" only emerged as distinct themes in 2025, roughly a year behind the global pattern.

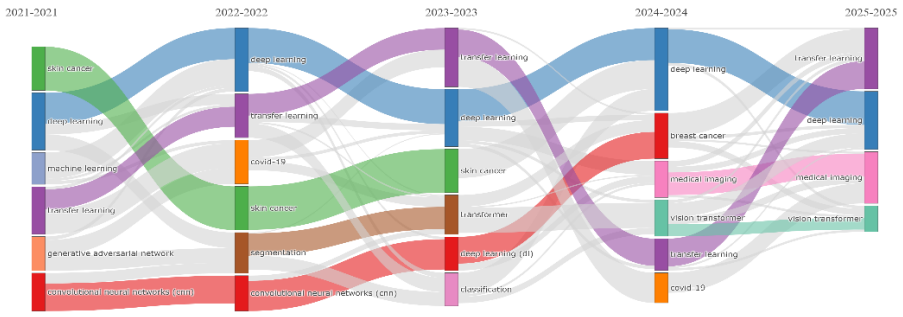


Figure 5. Global thematic evolution (Sankey diagram)

Figure 6 presents the corresponding thematic evolution for the Turkey-affiliated corpus.

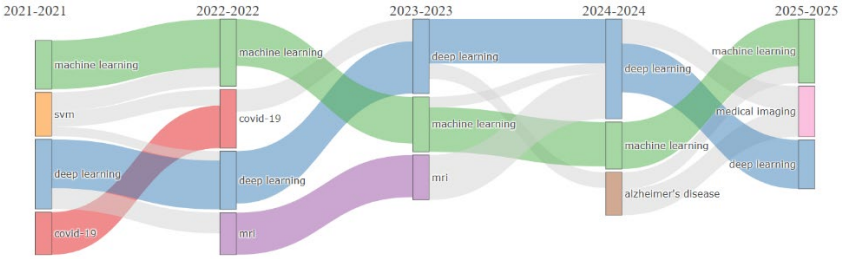


Figure 6. Turkey thematic evolution, 2021-2025, uniform annual bins (Sankey diagram)

4. Supplementary Ungrouped Keyword Trends

Grouping keywords into 23 themes averages out signals that are too small or too new to register at that scale. A supplementary, ungrouped pass ranked by 2024-to-2025 share growth (Tables 7 and 8) shows that globally, the steepest risers by volume were "deep learning," "medical imaging," and "vision transformer," but ranking instead by average citations per paper reorders this list: "vision transformers" (11.5 citations/paper) and "ensemble learning" (8.3) emerge as the highest-impact risers, while high-volume terms like "convolutional neural networks (cnn)" (2.7) rank lowest — indicating that keyword growth and citation impact identify different topics. Turkey's risers ran the other way: clinical decision support, risk prediction, and precision medicine. The clinical-decision-support overlap, arriving from different directions in each corpus, is the closest thing to a convergence signal here. However, the sample sizes are small enough to treat it as a future research lead rather than a settled finding. In the Turkish corpus, "clinical decision support" increased from 1 to 17 mentions between 2024 and 2025 but garnered only 126 citations across both years, suggesting that this signal should be read as an early-stage trend rather than an established finding.

Table 7: Citation-Weighted 2024→2025 Risers — Global

Keyword	2024→2025 f	Avg. citations/paper
Vision transformers	98→193	11.5
Vision transformer	278→414	8.5
Ensemble learning	44→85	8.3
Medical image analysis	108→228	6.9
Grad-CAM	13→86	6.6
Deep learning	1876→2084	5.8
Convolutional neural networks (CNN)	105→181	2.7

Table 8 reports the same ranking for the Turkey-affiliated corpus.

Table 8: Citation-Weighted 2024→2025 Risers — Turkey

Keyword	2024→2025 f	Total citations
Artificial intelligence	138→278	3,536
Machine learning	292→488	5,330
Clinical decision support	1→17	126
Large language models	6→19	434

Table 9 turns from these thematic rankings to the sampling-validity checks described in Section III.

Table 9: Sampling Validity Summary (Global Corpus, 5 Relevance-Ranked Slices)

Comparison	Spearman ρ	Jaccard J	χ^2 (p)
Consecutive slices (mean)	0.890	0.741	66.4–289.4 ($p < 0.001$)
Slice 1 → 5 (cumulative)	0.488	0.538	-

V. DISCUSSION

The dominance of CNNs and deep learning in the global dataset aligns with earlier bibliometric findings (Chen et al. [11]; Litjens et al. [14]). ViT's rapid rise after 2023 tracks the prediction Dosovitskiy et al. [8] made at the outset: with enough pretraining data, its ability to capture long-range dependencies makes it a genuine rival to CNN. Liu et al.'s Swin Transformer [22] reinforces the momentum, demonstrating higher segmentation accuracy on general-purpose vision benchmarks and gains that later motivated its adoption in medical imaging pipelines. The rise in interpretability terms, XAI, Grad-CAM, SHAP, LIME, reads as a direct response to the clinical-trust barrier flagged by Topol [23]; Selvaraju et al.'s Grad-CAM [17] built an early bridge by turning black-box decisions into visible heat maps.

Two structural weak points define the Turkish dataset before anything else does. U-Net occurs at 13.4 per 1,000 Turkish publications, compared with 32.3 globally. MRI shows a similar pattern, with 16.5 against 28.8. Both figures indicate that segmentation and neuroimaging capacity have not yet matured nationally. This is worth naming first because it forms the backdrop against which the more striking ratios below should be read. Turkey is not simply behind across the board. It is specifically thin on the infrastructure-heavy end of the field. The outsized share held by "machine learning" (2.01) and "artificial intelligence" (2.06) suggests many national studies still lean on traditional algorithms rather than deeper architectures. However, keyword data alone cannot distinguish a training gap from a slower architectural transition.

ChatGPT and related LLM terms move in opposite directions. When grouped into a single category (ChatGPT, GPT-4, LLM, large language model), 14 mentions appear globally, compared with 89 in the Turkish corpus, a gap that normalizes to 26.2-fold (χ^2=290.2, $p<0.001$). Grouping addresses the single-term fragility of the raw ChatGPT count: pooling ChatGPT with its generic-term variants absorbs the terminological split, where researchers outside Turkey may favor "large language model" over the branded name. The resulting ratio, still large but now statistically robust, indicates a genuine relative preference among Turkish authors for LLM- and ChatGPT-related terms. However, keyword counts alone cannot establish intent beyond that.

ViT tells a steadier story. Turkey's raw count of 51 against the global figure of 970 looks like a nineteen-fold gap, but once normalized by volume (10.5 versus 48.6 per 1,000), it narrows to about 4.6x, real under-adoption but far short of what the raw count implies. Radiomics presents a partial exception, since its 2.42x ratio is the highest in Table 2 after the grouped ChatGPT/LLM category's 26.2x outlier, despite Turkey's weak infrastructure

numbers on U-Net and MRI noted above, suggesting infrastructure-level and application-level capacity do not move together here. The fast-growing interest in ChatGPT in Turkey since 2024 also shows that the national research community is responding quickly to the generative AI wave. However, how far that interest actually feeds into medical imaging work remains a separate question for future full-text mining to answer.

The 4,000-record-per-year cutoff also introduces a systematic bias, since Scopus's relevance ranking rewards citation history and metadata completeness over recency. The Turkey dataset, processed in full without truncation, is largely free of this bias, which is why normalized densities are reported alongside raw counts throughout. Future work should draw on API access or an institutional license permitting full export.

Despite these limitations, the patterns identified here point to several practical directions. Encouraging ViT- and XAI-focused projects in Turkey would strengthen both the country's contribution to the global agenda and the transparency now expected of clinical decision-support systems under regulations such as the EU AI Act. The thin national footprint in segmentation and neuroimaging also points to opportunities for international collaboration, given Turkey's centralized health databases. Closer ties between medical and engineering departments, as Lin et al. [13] suggest, could help push national output further.

VI. CONCLUSION

This study examined AI-assisted medical imaging research from 2021 to 2025 through a comparative, volume-normalized bibliometric lens. Globally, CNN maintains its dominant position even as ViT, Swin Transformer, XAI, and Grad-CAM enter the thematic landscape after 2023, steering the field toward attention mechanisms and interpretability. Turkey's output still leans on machine learning and classification, with a grouped ChatGPT/LLM density 26.2 times the global rate ($p<0.001$) — a robust signal once single-term fragility is removed by grouping. Normalizing by corpus size reduces ViT's raw 19-fold gap to a 4.6-fold gap. This is a correction to scale, not to direction. Radiomics runs in the opposite direction entirely, with Turkey proportionally ahead of the global literature despite a weaker segmentation infrastructure. Clinical decision support is rising in both corpora, though from different starting points. It remains the only candidate signal worth tracking as the literature moves beyond its current small-sample stage. The relevance-ranked, year-stratified truncation was independently validated through a five-slice rank-stability analysis, supporting the conclusion that the working set reflects the field's dominant thematic structure. The caveats on low-citation and long-tail representation noted above still apply. Future work should extend coverage to WoS and PubMed, apply NLP-based full-text mining by methodology and clinical modality, and encourage Turkey-based research on ViT and XAI through TUBITAK-funded university-industry partnerships. A living bibliometric dashboard tracking individual keywords alongside the 23 thematic groups would also help, since signals like Mamba or clinical decision support surface at the keyword level well before they register thematically.

REFERENCES

[1]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[2]. M. A. Abdou, "Literature review: Efficient deep neural networks techniques for medical image analysis," Neural Comput. Appl., vol. 34, no. 8, pp. 5791–5812, 2022, doi: 10.1007/s00521-022-06960-9.

[3]. N. Guler and Z. Ben Hazem, "Semantic communication-based convolutional neural network for enhanced image classification," Franklin Open, vol. 9, p. 100192, 2024, doi: 10.1016/j.fraope.2024.100192.

[4]. N. Guler, Z. Ben Hazem, and A. Gunes, "Convolutional neural network-enabled semantic communication approach for image classification tasks," in Proc. 2024 Int. Conf. on Optimization, Information and Data Technologies (OIDT), 2024, pp. 1–8, doi: 10.1109/OIDT59407.2024.11082691.
[5]. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
[6]. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE CVPR, pp. 4700–4708, 2017, doi: 10.1109/CVPR.2017.243.
[7]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, pp. 6105–6114, 2019.
[8]. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR, 2021, doi: 10.48550/arXiv. 2010.11929.
[9]. M. Aria and C. Cuccurullo, "Bibliometrix: An R-tool for comprehensive science mapping analysis," J. Informetrics, vol. 11, no. 4, pp. 959–975, 2017, doi: 10.1016/j.joi.2017.08.007.
[10]. N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," J. Bus. Res., vol. 133, pp. 285–296, 2021, doi: 10.1016/j.jbusres.2021.04.070.
[11]. Y. Chen et al., "Trends and hotspots in research on medical images with deep learning: A bibliometric analysis from 2013 to 2023," Front. Artif. Intell., vol. 6, p. 1289669, 2023, doi: 10.3389/frai. 2023.1289669.
[12]. R. Wang et al., "Bibliometric analysis of the application of deep learning in cancer from 2015 to 2023," Cancer Imaging, vol. 24, no. 1, pp. 1–17, 2024, doi: 10.1186/s40644-024-00737-0.
[13]. M. Lin et al., "A bibliometric analysis of the advances of artificial intelligence in medicine," Front. Med., vol. 12, p. 1504428, 2025, doi: 10.3389/fmed.2025.1504428.
[14]. G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
[15]. D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," Annu. Rev. Biomed. Eng., vol. 19, pp. 221–248, 2017, doi: 10.1146/annurev-bioeng-071516-044442.
[16]. R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," Briefings Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018, doi: 10.1093/bib/bbx044.
[17]. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, pp. 618–626, 2017, doi: 10.1109/ICCV.2017.74.
[18]. E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," Expert Syst. Appl., vol. 167, p. 114161, 2021, doi: 10.1016/j.eswa.2020.114161.
[19]. N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, vol. 84, no. 2, pp. 523–538, 2010, doi: 10.1007/s11192-009-0146-3.
[20]. O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.
[21]. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017, doi: 10.1038/nature21056.
[22]. Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE ICCV, pp. 10012–10022, 2021, doi: 10.1109/ICCV48922.2021.00986.
[23]. E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nat. Med., vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

aBERT-Based Hybrid Models for Fake News Detection: A Systematic Review from 2021 to 2026

Dilara ÇELİK
Department of Computer Engineering
Gazi University
Ankara, Türkiye
26833301030@gazi.edu.tr
ORCID:0009-0005-4860-6460

Necaattin BARIŞÇI
Department of Computer Engineering
Gazi University Ankara, Türkiye
nbarisci@gazi.edu.tr
ORCID: 0000-0002-8762-5091

Yusuf SÖNMEZ
Department of Computer Engineering
Gazi University Ankara, Türkiye
ysonmez@gazi.edu.tr
ORCID: 0000-0002-9775-9835

Abstract
The rapid dissemination of fake news through social media threatens democratic processes and public trust. Although BERT provides a strong foundation for automated detection, its limitations in capturing linguistic, structural, and social dimensions motivate the development of hybrid architectures. This systematic review analyzes 18 peer-reviewed studies published between January 2021 and June 2026, identified through database searches conducted in May–June 2026. It classifies them into eight architectural categories: BERT+CNN, recurrent networks, gradient boosting, graph neural networks, stacking ensembles, multimodal models, GAN-based augmentation, and other approaches. Across studies reporting BERT-only baselines, hybrid models achieved higher performance, with accuracies ranging from 92% to 98%. BBC-FND reported the highest accuracy (98%), while BERT+CNN achieved the highest category mean (97.0%) across two studies. GNN integration provided approximately 2–3% additional improvement through social-contextual signals. Explainability remained limited: three studies (17%) used post hoc methods, six provided partial interpretability, and nine were largely black-box. Key gaps include dataset standardization, English-language dominance, insufficient statistical testing, and limited preparedness for LLM-generated fake news. The findings highlight the need for standardized, rigorous, interpretable, multilingual, and practically applicable BERT-based hybrid frameworks.

Keywords: fake news detection, BERT, hybrid deep learning, explainable artificial intelligence, natural language processing, systematic review

I. INTRODUCTION

The digital transformation of information production and consumption has enabled the rapid proliferation of fake news—content deliberately produced for political, economic, or ideological purposes [1] — and has increasingly framed it as information disorder [2]. Such content can spread up to six times faster than true news on social media [3], threatening democratic processes, social trust, and public health. Moreover, LLM-generated misinformation can be more difficult for humans and automated detectors to identify than semantically equivalent human-written misinformation [4]. At the same time, existing detection methods degrade in the presence of LLM-generated disinformation [5], highlighting the increasingly adversarial nature of fake news detection [6]. BERT's strong benchmark performance [7], enabled by the Transformer self-attention mechanism [8], established a promising foundation through bidirectional attention, pretrained representations, and fine-tuning flexibility. However, BERT alone has limitations in capturing local linguistic patterns and long-range dependencies, incorporating social propagation and metadata, and maintaining computational efficiency. These limitations have motivated hybrid architectures that combine BERT with

convolutional, recurrent, and graph neural networks, as well as gradient-boosting methods. Existing reviews address only specific aspects of this field: Alnabhan and Branco [9] examine deep learning-based detection without synthesizing comparative performance; Athira et al. [10] focus on explainability without systematically evaluating transformer-based hybrids; and Nasser et al. [11] review multimodal approaches without a systematic taxonomy of fusion strategies. Thus, an integrated review that combines architectural taxonomy, cross-category performance analysis, and explainability assessment remains lacking, with only three of the eighteen studies in our corpus incorporating explicit post hoc explainability mechanisms. Accordingly, this review classifies BERT-based hybrid architectures, compares their performance across metrics, datasets, and methodologies, examines explainability, maps methodological gaps, and identifies future research directions. Specifically, it investigates the extent to which hybrid models outperform plain-BERT baselines (RQ1), which architectural combinations achieve the highest accuracy, and how performance varies across datasets (RQ2), how SHAP/LIME integration affects performance and interpretability (RQ3), and which methodological gaps and future directions remain in the field (RQ4). The remainder of the paper is organized as follows: Section II presents the methodology; Section III presents the architectural taxonomy and the reviewed corpus; Sections IV–VI present the performance, dataset, and explainability analyses; Section VII presents the methodological gaps and research directions; and Section VIII presents the conclusions.

II. REVIEW METHODOLOGY

This study was conducted as a systematic literature review in accordance with the PRISMA 2020 guidelines [12]. Searches were performed in Google Scholar, IEEE Xplore, Elsevier ScienceDirect, SpringerLink, Nature Portfolio, MDPI, and Web of Science using three Boolean keyword groups: (i) fake news and disinformation detection, (ii) transformer-based language models, and (iii) hybrid and ensemble learning approaches. Keywords within each group were combined with OR, and groups with AND. A representative query was: ("fake news" OR "disinformation" OR "misinformation") AND ("BERT" OR "RoBERTa" OR "transformer") AND ("hybrid" OR "ensemble" OR "CNN" OR "LSTM" OR "graph neural network"). Database-specific syntax was adapted using title, abstract, and keyword fields where available, and full-text searching otherwise. The search covered peer-reviewed journal articles published between January 2021 and June 2026. Retrieved records were merged and duplicates manually removed by matching titles, authors, and DOIs. Title/abstract screening and full-text eligibility assessment were conducted by the first author, with ambiguous cases resolved through discussion among all authors. The protocol was not publicly registered (e.g., PROSPERO.) Studies were included if they (i) were published between 2021 and 2026, (ii) appeared in peer-reviewed journals, (iii) used a transformer-based language model as the primary text encoder, (iv) addressed fake news or disinformation detection, (v) incorporated at least one additional learning component into a hybrid architecture, and (vi) reported at least one quantitative metric, such as accuracy or F1-score. Studies without an accessible full English text were excluded. One methodological exception was admitted: Truică and Apostol [13], which does not propose an explicit hybrid fusion architecture but provides a systematic and transferable methodological contribution relevant to the review questions. The search and screening process is summarized in Fig. 1. The database and backward citation searches initially yielded 487 records, which were reduced to 18 included studies after screening and full-text assessment. Two studies required particular consideration: Hashmi et al. [14] was included because its transformer encoder is complemented by a FastText embedding layer, satisfying the hybrid framework. At the same time, Truică and Apostol [13] were retained as the methodological exception because they systematically compare embedding methods and report significance using the Wilcoxon signed-rank and McNemar tests.

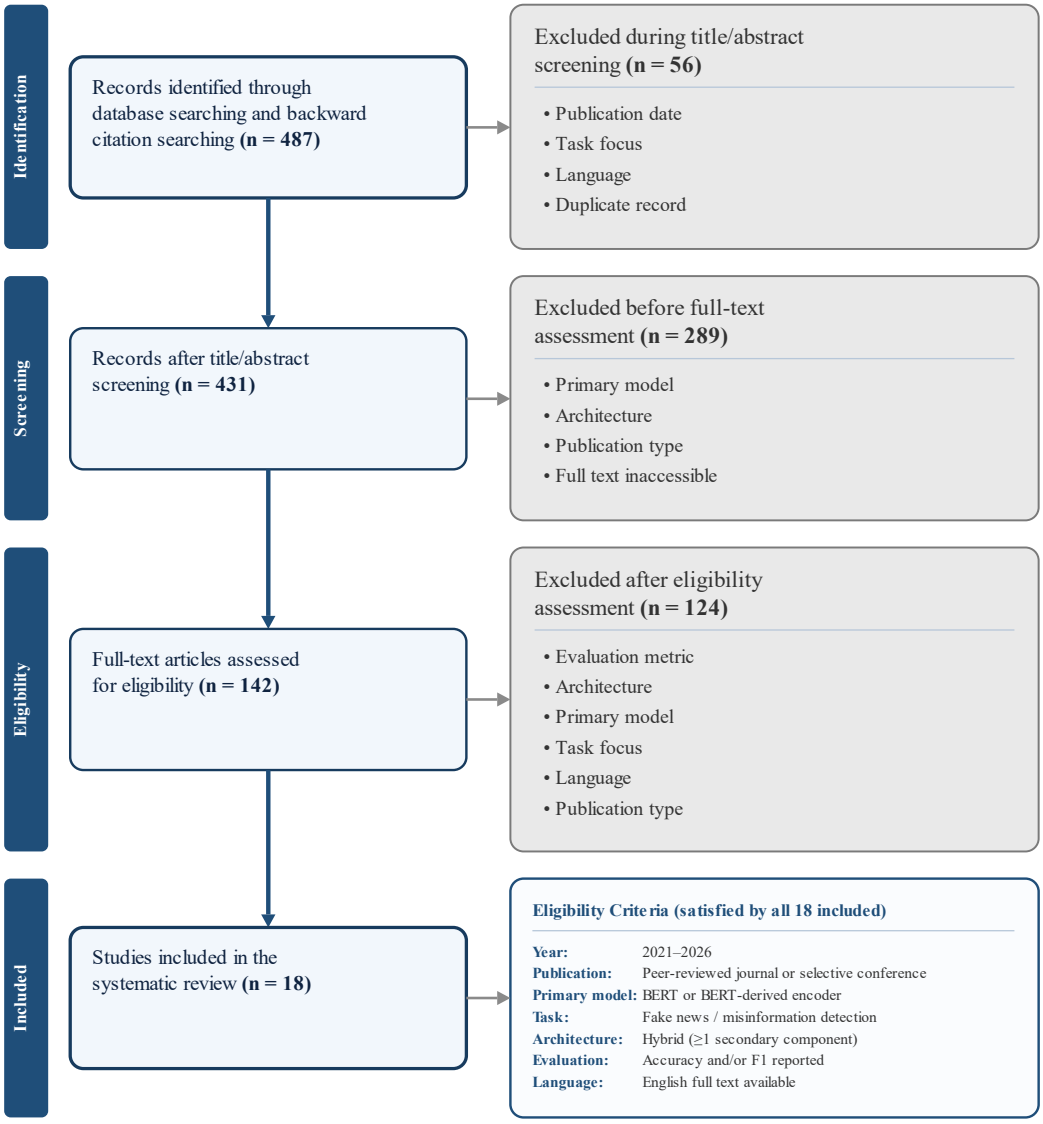


Figure 1. PRISMA-based flow diagram illustrating the literature search, study selection, and eligibility screening process.

III. ARCHITECTURAL TAXONOMY OF BERT-BASED HYBRIDS

As shown in Table 1, 12 of the 18 studies were published in JCR 2024 Q1 journals and five in Q2, while Human-Centric Intelligent Systems is not yet JCR-indexed but is classified as Scopus Q1. Graph neural network approaches, typically implemented using graph attention networks [15] or graph convolutional networks [16], constitute the largest category, with four studies, followed by stacking ensembles, with three.

BERT+CNN, recurrent neural networks, gradient boosting, multimodal, and other approaches each account for two studies, whereas GAN-based augmentation is represented by one study. Representative works include BERT–CNN hybrids [17], [18], multimodal capsule- and CNN–GRU-based models [19], [20], gradient-boosting hybrids [21], [22], a sentiment-aware BERT ensemble [23], and a GANBERT–BiLSTM ensemble [24]. CNNs [25] are the most frequently used complementary component, followed by recurrent layers (LSTM [26], GRU [27]) and gradient-boosting classifiers (XGBoost [28], LightGBM [29]). BERT-base serves as the primary backbone in most studies, while RoBERTa [30] is used in six studies requiring stronger semantic representations. Four main fusion strategies are identified: parallel concatenation, the most common approach; hierarchical sequential feeding, as in FakeStack [31], which sequentially combines BERT, CNN, and LSTM; stacking at the meta-learner level, as in BBC-FND [32], which achieves the highest reported accuracy (98%) through error-complementary predictions; and attention-based fusion, where cross- and co-attention mechanisms [33] provide approximately 1.5% improvement over late fusion. Several studies provide distinctive contributions. GETE [34] reduces computational cost by 60% with only a 0.3% performance decrease through GraphSAGE sampling and provides the only systematic multilingual evaluation. Kukkar and Kaur's adaptive ensemble [35] dynamically adjusts component weights based on entropy, topic, and text length and uniquely combines LIME and SHAP. The dual-stream model of Rama Moorthy et al. [36] jointly processes textual and social-graph streams using cross-attention, representing the most advanced attention-based fusion in the reviewed corpus.

Table 1. Overview of the 18 reviewed studies.

#	Study [ref]	Venue (JCR 2024 quartile)	Hybrid model	Dataset(s) (approx. N)	Validation	BERT-only baseline	Acc/ F1 (%)
1	Kaliyar et al., 2021 [17]	Multimedia Tools Appl. (Q1)	FakeBERT (BERT+CNN)	LIAR (~12.8k), ISOT (~44k), WELFake (~72k)	80/10/10	Reported (n/r)	≥97 / ≥97
2	Lin et al., 2022 [23]	Inf. Process. Manage. (Q1)	BERT ensemble + VADER	Fake-news set (N n/r), Webis Clickbait 2017 (~38k)	75/15/10	~91–92	94.7 / 94.7
3	Palani et al., 2022 [19]	Multimedia Tools Appl. (Q1)	CB-Fake (CapsNet+BERT)	MediaEval VMU, Fakeddit (~1M), Twitter multimodal (~15k)	80/10/10	~91–92	95 / 95
4	Essa et al., 2023 [21]	Complex Intell. Syst. (Q2)	BERT + LightGBM	ISOT (~44k), TI-CNN (~13k), FNC (~75k)	80/10/10 + oversampling	Reported (n/r)	96 / 96
5	Keya et al., 2023 [31]	PLoS ONE (Q2)	FakeStack (BERT-CNN-LSTM)	LIAR (~12.8k), ISOT (~44k), FNN GossipCop (~22k)+PolitiFact (~1k)	70/15/15	~93–94	97 / 97
6	Palani & Elango, 2023 [32]	Comput. Electr. Eng. (Q1)	BBC-FND (BERT+BiLSTM +CNN)	Kaggle FN (~20k), WELFake (~72k)	80/10/10 + SMOTE	~94–95	98 / 98

7	Verma et al., 2023 [18]	J. Ambient Intell. Human. Comput. (Q2)	MCred (BERT+CNN)	Kaggle FN (~20k)	80/10/10 + augment.	~95.9	97 / 97
8	Al-Quayed et al., 2024 [22]	IEEE Access (Q2)	Hybrid Transformer (BERT/RoBERTa +XGBoost)	LIAR, ISOT, FNN GossipCop+PolitiFact, custom	80/10/10 + class weighting	~93 ([CLS]+XGBoost)	95 / 95
9	Khalil et al., 2024 [20]	Human-Centric Intell. Syst. (n/a)*	Hybrid NN (BERT+CNN+GRU+image)	Kaggle FN + images (~12k with image)	80/10/10	~93–94	96 / 96
10	Truică et al., 2024 [37]	Knowl.-Based Syst. (Q1)	DANES (BERT+BiLSTM +GAT)	FNN GossipCop+PolitiFact, PHEME (~330), BuzzFeed (~1.6k)	80/10/10	~89–90	93 / 93
11	Hashmi et al., 2024 [14]	IEEE Access (Q2)	FastText + CNN + BiLSTM	Twitter FN (N n/r), ISOT (~44k); LIAR+WELFake	80/10/10	BERT hybrid ~95.5	93 / 93
12	Zhang et al., 2024 [33]	Pattern Recognit. Lett. (Q1)	GBCA (BERT+GCN, co-attention)	Twitter (~10–15k), PolitiFact (~1k), GossipCop (~22k)	80/10/10	~91–92	95 / 95
13	Kumar et al., 2025 [34]	Sci. Rep. (Q1)	GETE (BERT+RoBERTa+GAT)	FNN GossipCop+PolitiFact, PHEME, Multilingual (EN/DE/FR/ES)	80/10/10	~91–92	95 / 95
14	Rama Moorthy et al., 2025 [36]	Sci. Rep. (Q1)	Dual-Stream (BERT+GAT, cross-attn)	FNN GossipCop+PolitiFact, Twitter15 (~1.49k), Twitter16 (~818), PHEME	80/10/10	~91–92	96 / 96
15	Brinda et al., 2025 [24]	Knowl. Inf. Syst. (Q1)	EnsembleNet (GANBERT+BiLSTM)	ISOT (~44k), PolitiFact (~3–5k) + GAN synthetic	80/10/10 + imbalance exps.	Reported (n/r)	96 / 96
16	Kukkar & Kaur, 2025 [35]	Expert Syst. Appl. (Q1)	Adaptive ensemble + LIME/SHAP	LIAR, WELFake, ISOT, FNN GossipCop+PolitiFact	5-fold CV	Fixed-weight ens. ~94–94.5	96 / 96
17	Truică & Apostol, 2023 [13]	Mathematics (Q1)	Document embeddings (RoBERTa)	CLEF CheckThat! 2021 (~12–15k)	5-fold CV	Fine-tuned BERT 91.2	92 / 92
18	Xia et al., 2026 [38	Ann. Oper. Res. (Q1)	BERT-LSTM + SHAP	Custom social-media + engagement metadata (N n/r)	75/15/10	Reported (hybrid +2–3)	95 / 95

IV. PERFORMANCE SYNTHESIS

Reported accuracy ranges from 92% to 98%, with the highest value (98%) achieved by the stacking-ensemble BBC-FND model [32]. BERT+CNN (97.0%) and stacking ensembles (96.2%) have the highest category-level means, whereas "other approaches" have the lowest (92.5%). These values are descriptive only, as categories contain one to four studies and involve heterogeneous datasets, class distributions, preprocessing, train–test splits, and evaluation protocols; therefore, no pooled effect size, confidence interval, or between-architecture significance test is justified, and no claim of architectural superiority is made. Ablation results consistently support hybridization: every study reporting a plain-BERT baseline shows improved performance with the hybrid configuration. In FakeStack [31], accuracy increases from approximately 93–94% with BERT alone to 95–96% with pairwise combinations and 97% with the full BERT-CNN-LSTM architecture. GNN integration provides an additional 2–3% improvement through social-context signals, based on within-study ablations [33], [34], [36], [37] rather than a cross-study average; in DANES [37], for example, the social branch increases accuracy from approximately 89–90% to 93%. This advantage is constrained by the lack of propagation data for newly published news. Precision and recall were excluded because of inconsistent reporting, while accuracy and F1 were sufficiently available for synthesis. Reported accuracy values follow the original studies, using lower bounds where no point estimate was provided (≥97% for FakeBERT). Direct numerical comparisons should therefore be interpreted cautiously. Although no formal reporting bias instrument was applied, every study reporting a plain-BERT baseline favors the hybrid model, with no null or negative results. This uniformity suggests potential publication or reporting bias and indicates that the reported gains may represent an optimistic upper bound on hybridization benefits. More than 14 datasets were used, enabling evaluation across diverse content types and platforms but limiting direct comparison. ISOT was used in eight studies, followed by FakeNewsNet, whose GossipCop and PolitiFact subsets each appeared in six. FakeNewsNet uniquely combines textual and social-propagation information, although its approximately 1,000-sample PolitiFact subset may limit statistical reliability and generalizability. LIAR appeared in five studies, while WELFake and Kaggle Fake News appeared in four each; three studies used custom datasets, reducing reproducibility. Other datasets—including PHEME, Fakeddit, Webis Clickbait 2017, CLEF CheckThat! In 2021, Twitter15/Twitter16, BuzzFeed News, and MediaEval VMU were used less frequently. Social-propagation information was available only in FakeNewsNet, PHEME, and Twitter15/Twitter16; the remaining datasets primarily provided textual content and metadata. Because no standardized risk-of-bias instrument exists for machine-learning benchmark studies, dataset and reproducibility risks were assessed using a transparent Low/Medium/High rubric. Dataset risk reflected benchmark availability, while reproducibility risk reflected data accessibility and reporting of model configurations; explainability and statistical significance were assessed separately. Study-level results are presented in Table 2, providing a broader quality assessment than the journal quartile alone.

Table 2. Per-study risk-of-bias and reporting-quality assessment.

#	Study [ref]	Dataset risk	Reproducibility risk	Explainability	Significance testing	Notes
1	Kaliyar [17]	Low	Medium	None	Not reported	High compute cost
2	Lin [23]	Medium	High	None	Not reported	Multi-task (clickbait+FN)
3	Palani (CB-Fake) [19]	Medium (noisy labels)	Medium	None	Not reported	Requires image modality
4	Essa [21]	Low	Medium	Partial	Not reported	Two-stage pipeline

5	Keya (FakeStack) [31]	Low	Medium	None	Not reported	7-config ablation
6	Palani & Elango [32]	Low	Medium	None	Not reported	Highest accuracy (98%)
7	Verma (MCred) [18]	Medium	Medium	Partial	No (author's note of absence)	Text-only in practice
8	Al-Quayed [22]	High (custom)	High	Partial	Not reported	3-stage pipeline
9	Khalil [20]	Medium (small ~12k)	Medium	None	Not reported	Limited rigor
10	Truică (DANES) [37]	Medium (~1k PolitiFact)	Medium (social-data)	Partial	Not reported	Text + social context
11	Hashmi [14]	Medium (undocumented)	High	Full (LIME)	Not reported	Efficiency-focused
12	Zhang (GBCA) [33]	Medium (~1k PolitiFact)	Medium (social-data)	Partial	Not reported	8-config ablation
13	Kumar (GETE) [34]	Medium (~1k PolitiFact)	Medium (social-data)	None	Not reported	Multilingual eval.
14	Rama Moorthy [36]	Medium (rumor task)	Medium (API-dep.)	Partial	Not reported	The task partly differs
15	Brinda (EnsembleNet) [24]	Medium (synthetic)	Medium	None	Not reported	Synthetic representativeness
16	Kukkar & Kaur [35]	Low	Low (5-fold CV)	Full (LIME+SHAP)	5-fold CV (no formal test)	Adaptive ensemble
17	Truică & Apostol [13]	Medium (CLEF)	Low (5-fold CV)	None	Yes (Wilcoxon+ McNemar)	Few reporting significance
18	Xia [38]	High (custom)	High (non-benchmark)	Full (SHAP)	Not reported	Engagement metadata

V. INTERPRETABILITY OF HYBRID DETECTORS

Explainability is the most prominent methodological deficiency in the reviewed literature. Only three of the eighteen studies employ formal post-hoc explainability methods: one explicitly fuses explainable AI with a BERT–LSTM architecture [38], using SHAP [39] or LIME [40] as the underlying techniques; six provide partial interpretability through attention weights or feature importance; and nine remain entirely black-box. Cross-attention visualizations provide diagnostic insight but are classified as having partial interpretability because they do not constitute post hoc explanation methods. A consistent linguistic pattern emerges in the explainability studies [14], [35], [38]. SHAP and LIME identify uncertainty expressions such as "allegedly," "sources claim," and "reportedly" as strong indicators of fake news. In contrast, reference expressions such as "confirmed," "official statement," and "according to" are associated with genuine news. The reported 72%

overlap and 28% divergence between SHAP and LIME stem exclusively from the adaptive-ensemble study by Kukkar and Kaur [35], the only study that jointly applies both methods and measures their agreement; therefore, these values should not be generalized to the full corpus. This finding suggests that explainability methods can support both model transparency and validation of learned linguistic markers. The lower cross-study mean accuracy of explainability-integrating studies should not be interpreted as evidence that interpretability reduces performance. Within-study ablations in the three relevant studies [14], [35], [38] report comparable accuracy with and without the post-hoc explanation layer, with one study [38] reporting a marginal improvement attributed to the removal of noisy features. However, because these studies differ in datasets and backbone architectures, this represents a within-study observation rather than a controlled cross-study comparison. Thus, the reviewed evidence does not indicate a necessary trade-off between explainability and detection performance.

VI. METHODOLOGICAL LIMITATIONS AND RESEARCH GAPS

The reviewed studies demonstrate a clear evolutionary trajectory from BERT+CNN models in 2021 to multimodal and multi-task approaches (2022), stacking and gradient-boosting methods (2023), cross-attention and graph-based models (2024), multilingual graph and GAN-based approaches (2025), and explainability-integrated BERT-LSTM models (2026). Despite this progression, several methodological gaps remain. Dataset standardization and explainability are the most critical limitations. All 18 studies use heterogeneous datasets and evaluation protocols, while 15 lack genuine post-hoc XAI mechanisms. A unified benchmark combining ISOT, WELFake, and FakeNewsNet, along with standardized SHAP/LIME integration, is therefore recommended. The English-only constraint applies to 17 studies, requiring multilingual evaluation using mBERT or XLM-R. Social-data dependency affects four studies, suggesting the need for text-only fallback mechanisms. Other limitations include 512-token truncation across all 18 studies, which motivates long-context models such as Longformer, BigBird, or hierarchical BERT; insufficient statistical rigor in 16 studies, for which significance tests and bootstrap confidence intervals are recommended; and opaque custom datasets in three studies, highlighting the need for transparent collection and labeling protocols. None of the studies evaluates LLM-generated fake news or reports streaming/incremental inference. Given evidence that LLM-generated misinformation can evade existing detectors [4], [5], LLM-authored benchmark corpora and robustness testing should be prioritized [6], alongside streaming APIs and incremental learning for real-time detection. Finally, computational cost affects eight studies, motivating distillation, pruning, and GraphSAGE sampling. These gaps define the main priorities for more standardized, interpretable, multilingual, statistically rigorous, LLM-robust, and deployable BERT-based detection systems.

VII. CONCLUSION

This review systematically evaluated 18 peer-reviewed studies published between 2021 and 2026 across 8 architectural categories, presenting the current state, evolution, and limitations of BERT-based hybrid models for fake news detection within an integrated framework. The findings show that in all studies reporting a plain-BERT baseline, the hybrid configuration is superior, that accuracy values range from 92% to 98%, and that there are modest but meaningful differences among architectural categories. The most striking observation pointing to the field's immature aspects, however, lies in the explainability dimension: only three of the reviewed studies incorporate a genuine explainability mechanism, revealing that the transparency gap is one of the most critical shortcomings in the field. The modest contribution of graph neural network integration via social-context signals, in turn, suggests that structural signals have not yet been sufficiently exploited. The greatest challenge facing the field is moving beyond technical accuracy to develop systems that are explainable,

multilingual, deployable in real time, and statistically rigorous. This review aims to guide future work both by offering a comparative panorama of the current state and by identifying concrete research priorities.

REFERENCES

[1]. [1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017.
[2]. [2] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," Council of Europe, Strasbourg, France, Rep. DGI(2017)09, 2017.
[3]. [3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.
[4]. [4] C. Chen and K. Shu, "Can LLM-generated misinformation be detected?" in Proc. Int. Conf. Learn. Representations (ICLR), 2024.
[5]. [5] B. Jiang, Z. Tan, A. Nirmal, and H. Liu, "Disinformation detection: an evolving challenge in the age of LLMs," in Proc. SIAM Int. Conf. Data Mining (SDM), 2024, pp. 427–435, doi: 10.1137/1.9781611978032.50.
[6]. [6] C. Chen and K. Shu, "Combating misinformation in the age of LLMs: opportunities and challenges," AI Magazine, vol. 45, no. 3, pp. 354–368, 2024, doi: 10.1002/aaai. 12188.
[7]. [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[8]. [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
[9]. [9] M. Q. Alnabhan and P. Branco, "Fake news detection using deep learning: A systematic literature review," IEEE Access, vol. 12, pp. 114435–114459, 2024.
[10]. [10] A. B. Athira, S. D. Madhu Kumar, and A. M. Chacko, "A systematic survey on explainable AI applied to fake news detection," Eng. Appl. Artif. Intell., vol. 122, art. 106087, 2023.
[11]. [11] M. Nasser, N. I. Arshad, A. Ali, H. Alhussian, F. Saeed, A. Da'u, and I. Nafea, "A systematic review of multimodal fake news detection on social media using deep learning models," Results Eng., vol. 26, art. 104752, 2025.
[12]. [12] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, art. n71, 2021, doi: 10.1136/bmj.n71.
[13]. [13] C.-O. Truică and E.-S. Apostol, "It's all in the embedding! Fake news detection using document embeddings," Mathematics, vol. 11, no. 3, art. 508, 2023.
[14]. [14] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali, and M. Abomhara, "Advancing fake news detection: Hybrid deep learning with FastText and explainable AI," IEEE Access, vol. 12, pp. 44462–44480, 2024.
[15]. [15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. Int. Conf. Learn. Representations (ICLR), 2018.
[16]. [16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. Int. Conf. Learn. Representations (ICLR), 2017.
[17]. [17] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia Tools Appl., vol. 80, no. 8, pp. 11765–11788, 2021.
[18]. [18] P. K. Verma, P. Agrawal, V. Madaan, and R. Prodan, "MCred: Multi-modal message credibility for fake news detection using BERT and CNN," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 8, pp. 10617–10629, 2023.

[19]. [19] B. Palani, S. Elango, and V. Viswanathan K, "CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT," Multimedia Tools Appl., vol. 81, no. 4, pp. 5587–5620, 2022.
[20]. [20] A. Khalil, M. Jarrah, and M. Aldwairi, "Hybrid neural network models for detecting fake news articles," Human-Centric Intell. Syst., vol. 4, no. 1, pp. 136–146, 2024.
[21]. [21] E. Essa, K. Omar, and A. Alqahtani, "Fake news detection based on a hybrid BERT and LightGBM models," Complex Intell. Syst., vol. 9, no. 6, pp. 6581–6592, 2023.
[22]. [22] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun, and T. S. Alnusairi, "Optimizing fake news detection: A hybrid transformer-based model for enhanced performance," IEEE Access, vol. 12, pp. 160822–160834, 2024.
[23]. [23] S.-Y. Lin, Y.-C. Kung, and F.-Y. Leu, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis," Inf. Process. Manage, vol. 59, no. 3, art. 102872, 2022.
[24]. [24] B. M. Brinda, C. Rajan, and K. Geetha, "EnsembleNet: A novel deep ensemble learning model with GANBERT and BiLSTM for automated fake news detection," Knowl. Inf. Syst., vol. 67, no. 11, pp. 11015–11040, 2025.
[25]. [25] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 1746–1751.
[26]. [26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[27]. [27] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014, pp. 1724–1734.
[28]. [28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794.
[29]. [29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 3146–3154.
[30]. [30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, arXiv:1907.11692.
[31]. [31] A. J. Keya, H. H. Shajeeb, M. S. Rahman, and M. F. Mridha, "FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection," PLoS One, vol. 18, no. 12, art. e0294701, 2023.
[32]. [32] B. Palani and S. Elango, "BBC-FND: An ensemble of deep learning frameworks for textual fake news detection," Comput. Electr. Eng., vol. 110, art. 108866, 2023.
[33]. [33] Z. Zhang, Q. Lv, X. Jia, W. Yun, G. Miao, Z. Mao, and G. Wu, "GBCA: Graph convolution network and BERT combined with co-attention for fake news detection," Pattern Recognit. Lett., vol. 180, pp. 26–32, 2024.
[34]. [34] C. Kumar, M. Bansal, M. A. Khan, V. Kaushik, M. Arquam, and A. Alabdultif, "Graph-augmented transformer ensemble framework for robust and scalable fake news detection in social media ecosystems," Sci. Rep., vol. 15, art. 31653, 2025, doi: 10.1038/s41598-025-31653-3.
[35]. [35] A. Kukkar and G. Kaur, "AEC: A novel adaptive ensemble classifier with LIME and SHAP-based interpretability for fake news detection," Expert Syst. Appl., vol. 281, art. 127751, 2025.
[36]. [36] H. Rama Moorthy, N. J. Avinash, N. S. Krishnaraj Rao, K. R. Raghunandan, R. Dodmane, J. J. Blum, and L. A. Gabralla, "Dual stream graph augmented transformer model integrating BERT and GNNs for context aware fake news detection," Sci. Rep., vol. 15, no. 1, art. 25436, 2025.

[37]. [37] C.-O. Truică, E.-S. Apostol, and P. Karras, "DANES: Deep neural network ensemble architecture for social and textual context-aware fake news detection," Knowl.-Based Syst., vol. 294, art. 111715, 2024.
[38]. [38] H. Xia, N. Islam, D. Zhu, J. Z. Zhang, A. Behl, and M. Roohanifar, "A novel method of fusing BERT-LSTM and XAI for social media disinformation identification," Ann. Oper. Res., 2026, doi: 10.1007/s10479-026-07063-7.
[39]. [39] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.
[40]. [40] M. T. Ribeiro, S. Singh, and C. Guestrin, "' Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Using Long Short-Term Memory (LSTM) Algorithm to Prevent the Foreign Currency Speculation in Libyan Black Markets

Zakaria Suliman Zubi
Dept of Computer Science, Faculty of Information Technology
Sirte University
Sirte, Libya
zszubi@s.edu.ly
ORCID: 0000-0002-6062-1517

Zakariya A Alsharef
Dept of Computer Science, Faculty of Information Technology, Libyan Academy
Misurata, Libya
zakariyaalsharef @Gmail.com
ORCID: 0009-0005-2138-3228

Abdulbsit M Alajnaf
Dept of Computer Science, Higher Institute of Science and Technology
Misurata, Libya
basitegnaf@gmail.com
ORCID: 0009-0008-4297-4512

Abstract

The exchange rate of the US dollar against the Libyan dinar is a crucial and influential topic for the Central Bank of Libya. It positively or negatively affects oil and gas revenues, thereby affecting the Central Bank's ability to cover public institutions' expenses in the country, which in turn impacts the lives of Libyan citizens. Speculation in the black market on the official exchange rate of the US dollar against the Libyan dinar is considered a crime against Libyan citizens and is punishable by law. This study employs a Deep Learning framework based on Long Short-Term Memory (LSTM) to detect and prevent speculation in the exchange rate of the US dollar against the Libyan dinar. Based on this, we applied the Long Short-Term Memory (LSTM) algorithm on the historical dataset of the official exchange rate against the dollar to detect speculation in the US dollar against the Libyan dinar. The LSTM algorithm shows promising results by leveraging extensive financial data and economic indicators to detect irregularities and precisely forecast future exchange rate movements in the black market. The experimental results of this paper demonstrate high predictive accuracy, with a root mean squared error (RMSE) of no more than 0.0428 Libyan dinars. By analyzing historical data over a 30-day timeframe, the model projected a potential price rise to 1.3010 Libyan dinars, representing a positive difference of 0.067 from the last actual closing price.

Keywords: Machine Learning, Financial Time Series Forecasting, Central Bank of Libya (CBL), Currency Speculation, Long Short-Term Memory (LSTM).

I. INTRODUCTION

Exchange rate stability is a fundamental element for the economic and financial equilibrium of any country. This importance becomes more pronounced in economies undergoing transition or those facing structural and political obstacles. In the Libyan context, following prolonged periods of political and economic instability, the Libyan Dinar (LYD) has come under severe pressure, leading to intensified foreign-currency black-market activity and aggressive currency speculation. Economic reports indicate that the widening gap between the official and parallel rates in Libya has led to structural distortions in resource allocation and increased import costs [1]. These phenomena not only threaten the Central Bank of Libya's (CBL) efforts toward monetary stability but also undermine citizens' purchasing power, hinder economic growth, and heighten uncertainty in the business environment. The central issue of this research is that the escalation of speculation in the foreign currency black market poses a significant obstacle to the Libyan economy. Specifically, the problem lies in traditional models' inability to account for the sudden fluctuations that characterize the Libyan market. Studies have shown that non-economic variables play a crucial role in directing exchange rates in conflict environments

[2]. This necessitates the immediate development of innovative analytical tools capable of accurately forecasting exchange rates, thereby strengthening efforts to counter these phenomena. Relying on traditional methodologies in the face of contemporary complexities may be insufficient for capturing the intricate patterns and interconnected interactions that govern the parallel exchange market. The significance of this research lies in providing an advanced analytical solution based on the Long Short-Term Memory (LSTM) algorithm, a sophisticated machine learning technique aimed at forecasting the US Dollar (USD) exchange rate against the Libyan Dinar (LYD). This algorithm is characterized by its superior ability to process data with "long-range dependencies," making it outperform classical statistical models in financial forecasting [3]. By leveraging a historical dataset of exchange rates from the Central Bank of Libya and integrating key economic and political variables, this research aims to provide a robust tool to assist policymakers at the Central Bank and other financial institutions. This tool will enable them to gain greater insight into market dynamics, predict future fluctuations, and formulate more effective policies to address economic and financial challenges. The applied methodology of this paper relies on processing historical time-series data of the price variable representing the closing rate. The data is transformed into the [0, 1] range using Min-Max Normalization to ensure training efficiency. The technical architecture of the model is based on defining a sequence length (Timesteps) of 30 days based on hyper-parameter tuning results, with data reshaped into a three-dimensional array (S×T×F) to enable the LSTM network to capture sequential patterns, ultimately extracting the predicted value using a Dense layer and the Adam optimizer with a learning rate of 0.01. This study seeks to achieve several primary objectives, including analyzing the factors contributing to the spread of the foreign currency parallel market and exchange rate volatility in Libya. Additionally, it aims to evaluate the effectiveness of the LSTM algorithm in forecasting USD/LYD exchange rates and to propose practical recommendations based on forecast results to support policy initiatives aimed at combating currency speculation.

II. LITERATURE REVIEW

The process of forecasting foreign exchange rates constitutes a fundamental challenge in financial and economic research, given the complex and non-linear nature of foreign exchange time series. Previous studies in this domain aim to develop computational models that extract hidden patterns and temporal dependencies from historical data, thereby supporting investment decisions and monetary policy. The research literature has witnessed a qualitative shift from relying on traditional standard models to adopting machine learning techniques, which have demonstrated superior ability to handle the volatility of financial markets [4].

1. Foreign Currency Speculation

Foreign currency speculation is a financial activity that demands high forecasting accuracy, as it relies heavily on traders' and financial institutions' ability to anticipate future price movements. In this context, machine learning algorithms have proven their efficiency as advanced tools to support speculation decisions. The Long Short-Term Memory (LSTM) network model is the most prominent choice in this field, due to its architecture's ability to retain information over long periods. The study by [5] confirmed the effectiveness of LSTM networks in forecasting the exchange rate of the Egyptian Pound against the US Dollar. It showed clear superiority over the traditional (ARIMA) model in prediction accuracy and error reduction, underscoring the efficiency of these networks in forecasting exchange rates in emerging economies undergoing rapid structural change. To enhance predictive performance. For instance, the study by [6] reviewed a hybrid model called (ALFA), which integrates LSTM networks with an Attention Mechanism. This integration aims to assign different importance weights to input features, and the model demonstrated superiority over traditional benchmark models. Similarly, [7] developed a hybrid model combining the Gated Recurrent Unit (GRU) with LSTM networks, aiming to leverage GRU's learning speed and LSTM's ability to capture long-term temporal dependencies. The results confirmed

that this hybrid model achieved very high accuracy in forecasting the future closing prices of major currency pairs.

Furthermore, some studies have focused on integrating LSTM with standard statistical models to enhance their ability to capture the specific characteristics of financial data. The study by [8] reviewed a hybrid model combining the standard (GARCH) model with an LSTM network to estimate Value at Risk (VaR) in the Rwandan foreign exchange market. The study aimed to utilize GARCH's ability to capture Volatility Clustering, and the results revealed that the hybrid model achieved up to a 10% improvement in prediction accuracy compared to standalone models. In an advanced context, the integration of non-financial data represents a significant trend. The study by [9] developed an innovative model based on "Deep Multimodal Fusion" to forecast the exchange rate of the US Dollar against the Chinese Yuan. The (MF-LSTM) model integrates financial time series data with "investor sentiment" extracted from social media texts using the BERT model. The results demonstrated that combining textual sentiment with historical data achieved significantly higher predictive accuracy than models relying on a single data type.

Despite this notable progress, research still points to persistent fundamental challenges, most notably the need for Real-time data processing and the difficulty of interpreting complex model predictions within the context of dynamic market dynamics [4].

III. THE OBJECTIVES

The objectives of this paper are to develop a comprehensive data-driven framework for the Central Bank of Libya that utilizes data science analytics techniques to address key challenges faced by the institution, such as the following:

- Develop a data-driven forecasting framework for the Central Bank of Libya that utilizes deep learning techniques to enhance monetary decision-making and mitigate foreign currency speculation.
- Analyze exchange rate fluctuations and temporal market trends in the parallel market to provide an automated early-warning mechanism for monetary authorities.
- Evaluate the predictive performance of the Long Short-Term Memory (LSTM) algorithm against non-linear temporal dynamics using daily historical exchange rate data.
- Formulate optimized time-series forecasting configurations (evaluating sequence lengths and hyperparameters) to generate reliable short-term USD/LYD exchange rate predictions for strategic planning.

IV. PROBLEM FORMULATION

The core problem addressed in this paper is that the violent fluctuations in the Libyan dinar's exchange rate on the black market have become too complex for traditional statistical models to comprehend and predict. We may realize the dimensions of this issue clearly in the following points:

- The Price Gap and Speculation: The exchange rate gap fueled speculation, distorting the economy and increasing living costs.
- The Shortcomings of Classic Models: Classic financial models fail in the Libyan market; they assume order where there is only high volatility, 'noise', and unpredictable price jumps.

- The Impact of Sudden Structural Shocks Standard analytical tools struggle to capture sudden, non-linear market shocks and severe exogenous events that cause abrupt regime shifts in parallel exchange rate dynamics within short timeframes.
- The "Long Memory" Challenge: Current tools fail to capture 'Long Memory'—the complex impact of past shocks on today's exchange rates—which requires specialized temporal analysis.

V. CONTRIBUTIONS

This paper builds qualitative scientific contributions to the use of advanced data analytics to predict exchange rate fluctuations, with a focus on the role of the Central Bank of Libya in reducing speculation and achieving financial stability:

- Algorithmic Innovations: Foreign Currency Market Trend Prediction: We apply Long Short-Term Memory (LSTM) networks to predict USD/LYD exchange rates, capturing complex temporal patterns in historical currency data.
- Supporting decision-making to combat speculation: The goal is to provide a practical solution that offers decision-makers at the central bank an early warning tool.
- Empirical Architecture Optimization: We evaluate sequential time-series configurations (establishing an optimal sequence length of 30 daily timesteps paired with hyperparameter tuning), proving the superiority of deep learning over classical statistical baselines in highly volatile environments.

VI. DATASET

1. Historical Dataset for USD/LYD Exchange Rates (2010–2020)

The dataset provided by the Central Bank of Libya contains 3484 records of daily historical exchange rates of the Libyan Dinar (LYD) against the US Dollar (USD) in the legal market (not the black market) from January 1, 2010, to December 31, 2020. The dataset includes the following columns:

- Date: The date of the exchange rate.
- Price: The closing exchange rate of the LYD to USD on the given date.
- Open: The opening exchange rate of the LYD to USD on the given date.
- High: The highest exchange rate of the LYD to USD on the given date.
- Low: The lowest exchange rate of the LYD to USD on the given date.
- Vol.: The volume of the LYD to USD exchange on the given date.
- Change %: The percentage change in the exchange rate from the previous day.

This dataset was used to train an LSTM (Long Short-Term Memory) model to predict future trends in the foreign currency market. The LSTM model can learn patterns and relationships between historical exchange rates and use this information to predict future exchange rates. By analyzing predicted trends, investors can make informed decisions about when to buy or sell the Libyan Dinar against the US Dollar in the foreign currency market (black market).

1	Date	Price	Open	High	Low	Vol.	Change %
2	12/31/2020	1.3332	1.3378	1.3392	1.3341	0.02K	-0.08%
3	12/30/2020	1.3343	1.34	1.3405	1.3313	0.02K	-0.16%
4	12/29/2020	1.3365	1.3396	1.3405	1.3364	0.02K	0.50%
5	12/28/2020	1.3299	1.3427	1.3427	1.3363	0.02K	-0.69%
6	12/25/2020	1.3392	1.3427	1.3427	1.3427		0.00%
7	12/24/2020	1.3392	1.3427	1.3433	1.3409	0.00K	0.00%
8	12/23/2020	1.3392	1.3412	1.3427	1.3351	0.02K	0.11%
9	12/22/2020	1.3377	1.3388	1.3413	1.3345	0.02K	0.18%
10	12/21/2020	1.3353	1.3393	1.3401	1.3342	0.03K	0.00%
11	12/20/2020	1.3353	1.3386	1.3386	1.3386		-0.04%
12	12/19/2020	1.3358	1.3393	1.3393	1.3393		0.00%
13	12/18/2020	1.3358	1.3392	1.3398	1.3341	0.02K	0.01%
14	12/17/2020	1.3357	1.3431	1.3431	1.3391	0.02K	-0.29%
15	12/16/2020	1.3396	1.3454	1.3454	1.3429	0.02K	-0.17%
16	12/15/2020	1.3419	1.3452	1.3462	1.342	0.02K	0.01%
17	12/14/2020	1.3417	1.3473	1.3481	1.3428	0.02K	-0.16%
18	12/12/2020	1.3438	1.3473	1.3473	1.3473		0.00%
19	12/11/2020	1.3438	1.3472	1.3479	1.3451	0.02K	0.01%
20	12/10/2020	1.3437	1.3404	1.3477	1.3401	0.02K	0.22%
21	12/9/2020	1.3408	1.3457	1.3463	1.3418	0.02K	-0.10%
22	12/8/2020	1.3422	1.3452	1.3463	1.3422	0.02K	0.04%
23	12/7/2020	1.3417	1.3466	1.3469	1.3433	0.02K	0.01%

Figure 1. Illustrates a sample of the historical daily exchange rates of Libyan Dinar (LYD) against the US Dollar (USD).

VII. METHODS AND ALGORITHMS

This section outlines the specific method and algorithm to be employed in the paper, including detailed descriptions, mathematical formulations, and illustrative examples using historical data derived from the Libyan financial sector.

- Combating Foreign Currency Speculation: Foreign Currency Market Trend Prediction using LSTM (Long Short-Term Memory)

LSTM is a type of recurrent neural network (RNN) designed to learn long-term dependencies in sequential data, making it suitable for predicting currency exchange rates[10].

- Data Preprocessing: To ensure efficient performance of the neural network, a set of steps was applied to the time variable price, which represents the closing price.
- Normalization: Before the data was entered into the model, it was converted to a range of [0, 1] using a Min-Max Normalization formula to avoid large values dominating the training process and to accelerate the convergence of the optimization algorithm.

$$X_{scaled} = (X - X_{min})/(X_{max} - X_{min}) \quad (1)$$

Where X is the original value of the price, and X_{min} and X_{max} are the minimum and maximum prices in the entire dataset.

- Time Series Sequencing: LSTM models rely on sequential input to learn time patterns. Based on the results of setting the characteristics, which showed that 30 days achieved the best accuracy, the length of the time-step memory (T) was set to 30 days.

The dataset was reshaped so that the model inputs are a three-dimensional (S×T×F) matrix where:

- S: Number of training samples.
- T: Length of time sequence (memory), T = 30 days.
- F: Number of features, F = 1 (where price was the sole consideration in the final model).

The relationship between inputs and outputs is defined as follows:

$$\{X_t\} = \{P_{t-29}, P_{t-28}, \dots, P_{t-1}, P_t\} \rightarrow Y_{t+1} = P_{t+1} \qquad (2)$$

Where X_t is the input vector, and Y_{t+1} is the expected closing price for the next day.

- LSTM (LSTM Network Architecture): A sequential model with a unilateral structure for the LSTM network was used for price prediction, and the optimal structure was determined based on experience:

Table 1 .LSTM Network Architecture

Layer	Type	Units	Function
1	LSTM	50 Units	The primary processing layer accepts input sequences of size (30×1).
2	Dropout	Rate: 0.2	Prevents **Overfitting** by randomly dropping 20% of the neurons' connections during training.

3	Dense	1 Unit	The output layer that generates the final prediction value (P_{t+1}).

- LSTM (The Core Mechanism): The LSTM (Long Short-Term Memory) cell features a gate mechanism that allows it to control the flow of information, enabling it to store data over long periods:
- Forget Gate (ft): Specifies the information to be discarded from the previous cell state (Ct−1).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

- Input Gate (it): Specifies which new values will be updated in the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \quad (4)$$

- Cell State Update (Cell State, Ct): The new cell state is updated based on forgetting and new input.

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{c_t}, \quad (5)$$

- Output Gate (ot): Specifies the cell state portion to be output as a hidden state (ht).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

Where σ is a sigmoid function, and W and b are the weight and bias matrices.

- Training and Optimization: The model was trained using settings that proved effective in achieving the lowest RMSE:
- Loss Function: The Mean Squared Error (MSE) function was used to measure the difference between actual and expected values. The training objective is to minimize this loss.

$$\text{MSE} = (1/N)\sum_{i=0}^{n}(y_i - \hat{y_i})^2, \quad (7)$$

Where N is the number of samples, y_i is the actual value, and yi is the expected value.

- Optimizer: The Adam algorithm was chosen because it is effective in deep learning applications and works well in handling non-stationary data.

  Learning Rate: 0.01

  Number of Courses (Epochs): 50

- Evaluation Metric: The model's performance on the test set was evaluated using the Root Mean Squared Error (RMSE) metric, which provides the error value in the original unit (Libyan Dinar LYD).

$$\text{RMSE} = \sqrt{\left((1/N)\sum_{i=0}^{N}(\text{y}_\text{i} - \hat{y}_i)^2\right)}, \qquad (8)$$

VIII. IMPLEMENTATION

In this stage, a phase of building and improving a forecasting model for the exchange rate of the US dollar against the Libyan dinar (USD/LYD) using long-term short-term memory networks (LSTM) in the "RStudio" environment, based on the "Keras" and "TensorFlow" libraries:

1. Determine the path and upload the data

This stage involves setting up the software work environment by linking the R platform to the Python engine via the reticulate package to ensure compatibility with TensorFlow libraries, followed by the process of importing historical data from an external file using the read_delim function, which allows specifying the structural parameters of the file (such as the type of divider) to ensure that the data is properly organized in preparation for processing it.

2. Data Pre-processing

A historical dataset of the US dollar-Libyan dinar exchange rate was imported and converted to numeric values to ensure the stability of the neural network and accelerate its convergence. Min-max scaling was applied to compress values within the range [0],[1]. The data was also converted from a one-dimensional time series to a supervised learning format using a 30-day sliding window as input (X) to predict the price for the following day (Y).

```
reticulate::use_python(
  "C:/Users/ALNOOR/Documents/.virtualenvs/r-tensorflow/Scripts/python.exe",
  required = TRUE)

# Load the dataset
data <- read_delim(
  file.choose(),
  delim = ";",
  col_names = TRUE
)
```

Figure 2. RStudio script for environment setup and dataset ingestion

3. Time-Series Windowing

Convert the data into a supervised learning format, a 30-day sliding window (X) was selected via hyperparameter tuning to predict the 31st-day price (Y).

```
# Reshaping data into sliding windows (30 days look-back)
create_sequences <- function(data, timesteps) {
  X <- list(); Y <- list()
  for (i in 1:(length(data) - timesteps)) {
    X[[i]] <- data[i:(i + timesteps - 1)] # Input features
    Y[[i]] <- data[i + timesteps]         # Target label
  }
  X_array <- array(unlist(X), dim = c(length(X), timesteps, 1))
  return(list(X = X_array, Y = Y_array))
```

Figure 3. Configuring the R-to-Python interface using reticulate and importing the USD/LYD

4. Model Architecture

The study employed a Sequential model featuring a 50-unit Long Short-Term Memory (LSTM) layer, followed by a 20% Dropout layer to mitigate overfitting. A single-unit Dense layer was used to produce the price output. The model was compiled with the Adam optimizer and Mean Squared Error (MSE) loss, and trained for 100 epochs with a batch size of 32.

```
# Compiling and training the model
model_final %>% compile(
  loss = "mean_squared_error",
  optimizer = optimizer_adam(learning_rate = 0.01)
)

model_final %>% fit(x = X_train, y = Y_train, epochs = 50, batch_size = 32)
```

Figure 4. Model compilation using Adam optimizer and Mean Squared Error (MSE)

5. Training and improvement

The model was compiled with the Adam optimizer and a learning rate of 0.01, and trained with Mean Squared Error (MSE) as the loss function. Training was executed over 50 epochs with a batch size of 32, ensuring efficient convergence and optimal temporal pattern extraction.

```
# Defining LSTM model architecture
model_final <- keras_model_sequential() %>%
  layer_lstm(units = 50, input_shape = c(timesteps, 1)) %>%
  layer_dropout(rate = 0.2) %>%
  layer_dense(units = 1)
```

Figure 5. Defining the LSTM model architecture with a Dropout layer to prevent overfitting and a Dense output layer

6. Final price prediction

The optimized model (30-day memory, RMSE = 0.0428 LYD) was used to predict the expected price after the last day in the dataset:

Table 2. Prediction Results (LYD)

Metric	Value (LYD)
Last Actual Closing Price	1.2340
Predicted Price for the Next Day	1.3010

IX. EXPERIMENT AND RESULT

The final evaluation phase revealed the proposed model's ability to predict price trends with high statistical accuracy; the model achieved an RMSE of 0.0428 Libyan dinars. Based on a temporal analysis of historical data, the out-of-sample forecast indicates a potential upward trend in the US dollar's exchange rate against the Libyan dinar. Table 2 compares the last actual closing price with the expected value, clearly showing a positive difference that points to bullish expectations.

Table 3. LSTM Model Architecture

Metric	Value (LYD)	Notes / Observations
Last Actual Closing Price	1.2340 LYD	The last known price provided to the model.
Predicted Next-Day Price	1.3010 LYD	The final output/forecast generated by the model.
Difference (Forecast - Actual)	+0.067 LYD	The model predicts a relatively significant price increase.

This projected increase of 0.067 LYD reflects the LSTM algorithm's response to dynamic patterns detected within the time-memory window, accounting for the observed mean margin of error. These results demonstrate the model's effectiveness in providing proactive indicators of exchange rate movements, thus offering a valuable analytical tool for decision-makers during periods of monetary volatility.

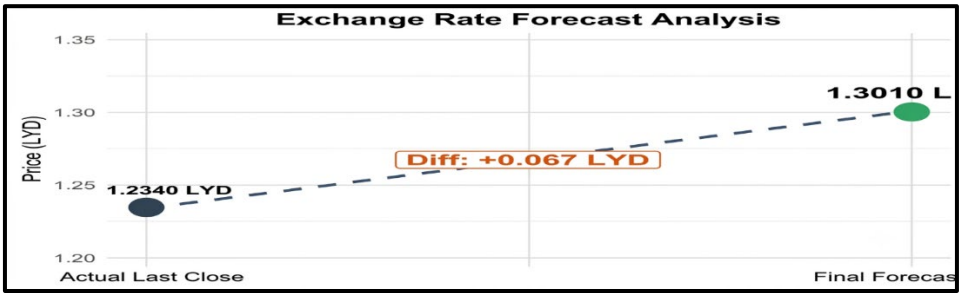


Figure 6. Comparison between the last actual closing price and the predicted price

Figure 6 shows the model's predictive output, comparing the last recorded actual closing price (1.2340 LYD) with the predicted price for the following day (1.3010 LYD). The upward trajectory illustrated between the two points highlights the LSTM model's capability to extrapolate temporal patterns and forecast future price movements. This visualization is analytically significant because it identifies the expected direction of the trend, helping decision-makers anticipate potential exchange rate fluctuations.

X. PERFORMANCE COMPARISON AND ANALYSIS

To determine the optimal architecture, four distinct LSTM model scenarios were evaluated using Root Mean Squared Error (RMSE) as the primary criterion. Table 4 presents the empirical performance comparison across different input configurations and hyperparameters.

Table 4. Model Performance Comparison

Model Type	Features (Inputs)	Memory (Timesteps)	Units	RMSE (LYD)	Evaluation
Original (Univariate)	Price only	10 Days	50	0.0494	Baseline Performance
Multivariate	Features (P, O, H, L, C)	10 Days	50	0.0904	Weak Performance (Overfitting)
Tuned	Price only	10 Days	100	0.1116	Weak Performance (Low Learning Rate)
Final & Optimized	Price only	30 Days	50	0.0428	Best Performance

While the baseline setup (Model 1) yielded acceptable results, performance degraded in Models 2 and 3 due to overfitting and suboptimal learning-rate adjustments. Notably, Model 4 achieved the highest precision with an RMSE of 0.0428 LYD, confirming that extending the temporal memory depth to 30 days was decisive in minimizing prediction errors.

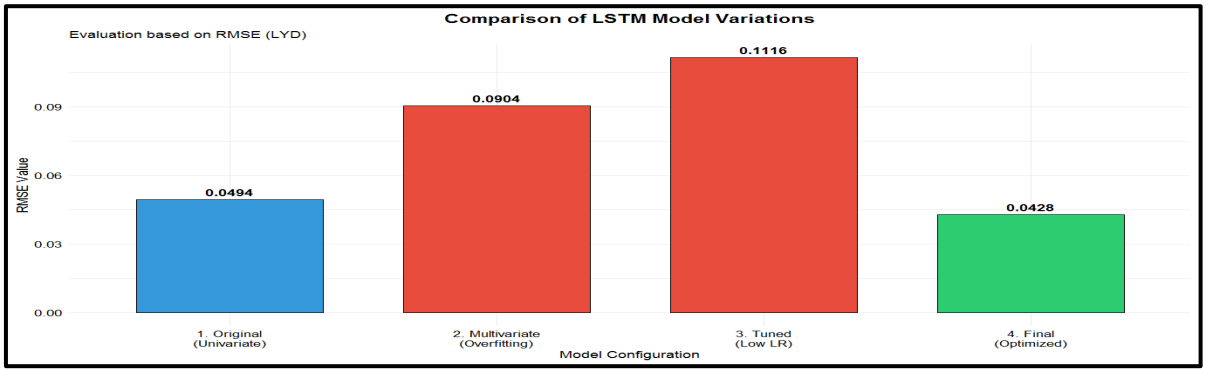


Figure 7. Performance comparison of LSTM model variants based on RMSE values

As illustrated in Figure 7, the comparative analysis highlights the model's sensitivity to learning rate and memory length, with the distinct drop in the fourth column demonstrating Model 4's superior efficiency in reducing predictive deviations.

XI. CONCLUSION

This paper presented a short-term predictive framework based on a Long Short-Term Memory (LSTM) neural network architecture to analyze and forecast USD/LYD exchange rate trends using an empirical dataset from the Central Bank of Libya. The experimental evaluations demonstrated the model's strong capability to capture complex non-linear financial time-series dynamics, achieving high predictive accuracy with a Root Mean Squared Error (RMSE) of 0.0428 LYD. By analyzing historical daily sequences with a 30-day temporal window, the model provided out-of-sample projections indicating an upward trend to 1.3010 LYD, marking a positive variation of +0.067 LYD over the last recorded closing price. Furthermore, the expected speculative trend indicator of 5.43% was identified by subtracting the previous closing price of 1.2340 LYD from the predicted price of 1.3010 LYD, then dividing the result by the previous closing price and multiplying by 100, representing the speculator's expected yield and serving as an early warning signal for the Central Bank. These empirical findings confirm that deep learning techniques, particularly optimized LSTM networks, serve as reliable predictive tools for financial time series, empowering central bank decision-makers and financial analysts with proactive analytical insights to mitigate market uncertainty and curb parallel market currency speculation.

REFERENCES

[1] International Monetary Fund, "Libya: 2023 Article IV Consultation—Press Release; Staff Report; and Statement by the Executive Director for Libya," IMF, Washington, DC, USA, IMF Country Rep. 23/201, Jun. 2023.

[2] ESCWA and GIZ, "A Preliminary Study on the Economy in Libya: Reality, Challenges and Prospects," United Nations, Beirut, Lebanon, Rep. E/ESCWA/CL6.GCP/2020/TP.3, 2020.

[3] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," FAU, Erlangen-Nuremberg, Germany, Discussion Paper 11/2017, 2017.

[4] S. Giantsidi and C. Tarantola, "Deep learning for financial forecasting: A review of recent trends," Int. Rev. Econ. Finance, vol. 104, p. 104719, Jan. 2025, doi: 10.1016/j.iref.2024.104719.

[5] M. M. N. Kamouh, "Forecasting Exchange Rates Using Artificial Neural Networks: An Applied Study on the Arab Republic of Egypt," Arab J. Admin., vol. 46, no. 6, Oct. 2024.

[6] S. Ghahremani and U. T. Nguyen, "Prediction of foreign currency exchange rates using an attention-based long short-term memory network," Mach. Learn. Appl., vol. 20, p. 100648, Mar. 2025, doi: 10.1016/j.mlwa.2024.100648.

[7] M. S. Islam and E. Hossain, "Foreign exchange currency rate prediction using a GRU-LSTM hybrid network," Soft Comput. Lett., vol. 3, p. 100009, Dec. 2021, doi: 10.1016/j.socl.2020.100009.

[8] E. Nsengiyumva, J. K. Mung'atu, and C. Ruranga, "Hybrid GARCH-LSTM Forecasting for Foreign Exchange Risk," FinTech, vol. 4, no. 1, pp. 22–38, Mar. 2025, doi: 10.3390/fintech4010002.

[9] E. Windsor and W. Cao, "Improving exchange rate forecasting via a new deep multimodal fusion model," Appl. Intell., vol. 52, pp. 12160–12177, Jul. 2022, doi: 10.1007/s10489-022-03342-5.

[10] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," Neural Networks, vol. 1, pp. 339–356, 1988.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Towards Multimodal Artificial Intelligence (OMEAI) Framework for Early Detection of Postnatal Depression

Gleslie Sybil Ann A. Maraon
Department of Computer Science,
Faculty of Science and Technology
Middlesex University
London, United Kingdom
g.maraon@mdx.ac.uk
ORCID:0009-0001-8972-0113

Xiaohong W. Gao
Department of Computer Science,
Faculty of Science and Technology
Middlesex University
London, United Kingdom
x.gao@mdx.ac.uk
ORCID:0000-0002-8103-6624

Abstract
Postnatal depression (PND) remains a significant global public health concern, with many cases going undetected and adversely affecting the well-being of both mothers and their infants. This study presents an enhanced multimodal explainable artificial intelligence (OMEAI) framework for early postnatal depression screening by integrating facial features, vocal characteristics, textual sentiment, and Edinburgh Postnatal Depression Scale (EPDS) scores. The framework incorporates multimodal fusion, explainable artificial intelligence techniques, reliability & validation, and SOAP-based clinical reporting within a unified healthcare-oriented architecture. The feasibility of the framework was evaluated using the Extended DAIC-WOZ dataset for speech and language analysis, a dataset of 100 postnatal facial images, and a multimodal postnatal case integrating facial, vocal, textual, and psychometric information. Comparative analysis against single-modality baseline models demonstrated improved interpretability and classification performance, with the receiver operating characteristic area under the curve (ROC-AUC) increasing from 0.84 to 0.92. The integration of explainability and reliability components enhanced transparency and facilitated clinically meaningful interpretation of model outputs. The findings demonstrate the feasibility and potential of the OMEAI framework to support early postnatal depression screening and decision-making. The framework may contribute to the development of intelligent digital health technologies and maternal healthcare screening support systems. Future work will focus on validation across larger, more diverse postnatal populations and real-world clinical settings to support translation into clinical practice.

Keywords: Multimodal Biomarkers, Postnatal Depression, Postnatal, Clinical Decision Support, Reliability, Explainability, Clinical Interpretability, SOAP Reporting, McNemar, DAIC WOZ, Mental Health, Psychosocial

I. INTRODUCTION

Women's health during the postnatal period is an important public health concern, particularly because postnatal depression (PND) can affect both maternal well-being and child development. Over the past decade, PND prevalence has increased globally, with reported rates ranging from 10–20% in high-income countries such as the UK and up to 30% in Southeast Asia, raising further concerns during and after the COVID-19 period [1], [2]. Reports from the Royal College of Psychiatrists UK highlighted that thousands of mothers in England may experience postnatal depression annually [3]. Despite the widespread use of screening tools such as the Edinburgh Postnatal Depression Scale (EPDS), challenges remain because current methods rely heavily on self-reported responses and may not fully capture behavioral or emotional changes [4]. These limitations highlight the need for more accessible and reliable approaches for early postnatal clinical screening. Recent advances in artificial intelligence (AI) have enabled the use of behavioral signals, such as facial expressions,

speech patterns, and language use, for depression-related assessment [5-10]. Speech datasets such as DAIC-WOZ, along with computer vision and natural language processing techniques, have supported the identification of emotional and cognitive indicators associated with depressive symptoms. Explainable AI approaches, including SHAP and Grad-Cam, have also been introduced to improve transparency and help clinicians understand how AI systems generate predictions [11-13]. Structured documentation approaches, such as the SOAP (Subjective, Objective, Assessment, and Plan) format, have been shown to improve the completeness and organization of clinical records. Their use can support clearer communication and more interpretable reporting, making them suitable for integrating AI-assisted findings into routine clinical practice.[14]

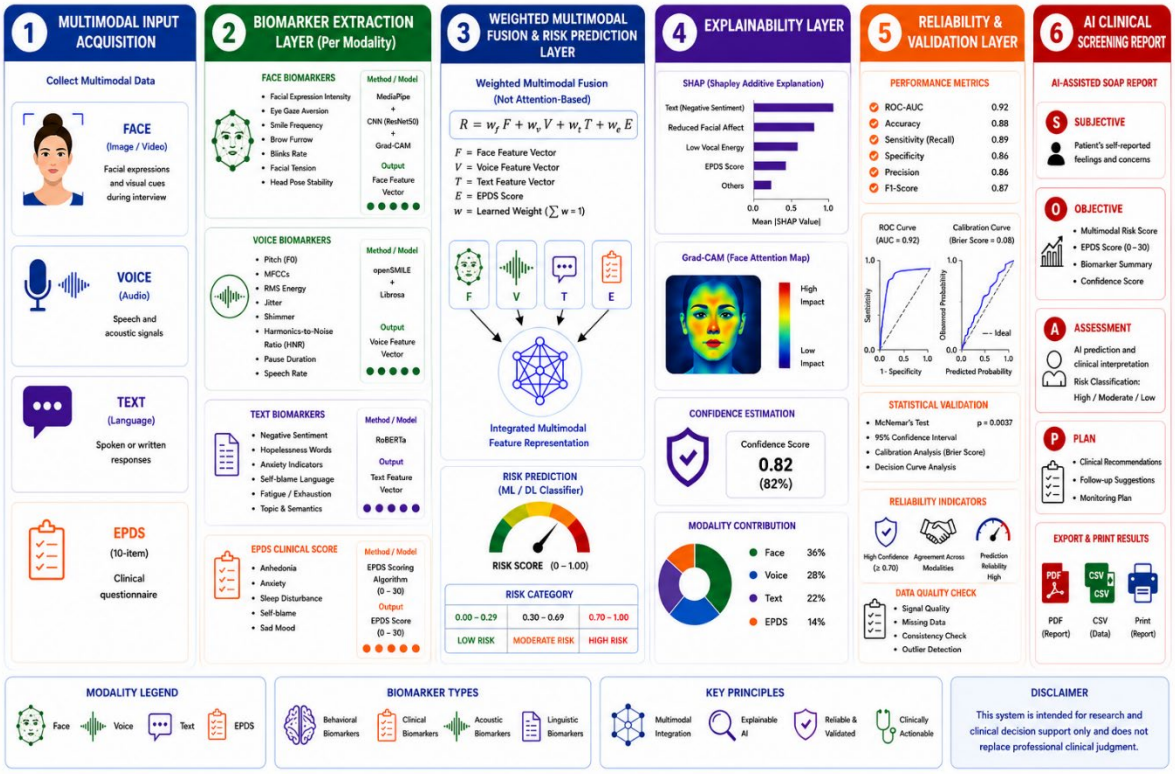


Figure 1. Proposed reliability and explainability-oriented multimodal AI framework integrating facial, vocal, textual, and psychometric biomarkers with explainable AI, statistical reliability validation, and SOAP-based clinical decision support for early detection of postnatal depression. However, many existing studies focus on single modalities or loosely integrate multimodal systems and often operate as black-box models with limited clinical interpretability [15-21]. In addition, psychometric tools such as the EPDS are commonly used independently based on text rather than integrated into AI-based predictive frameworks. These limitations highlight the need for more transparent and reliable multimodal approaches for postnatal depression screening.
In this study, a framework integrating multiple information modalities is proposed. As illustrated in Figure 1, the system embeds multi-modal information, including facial expressions (images), voices (audio), and questionnaires (text).

II. METHODOLOGY

This study developed a reliability- and explainability-oriented multimodal artificial intelligence (OMEAI) prototype framework (Figure 1) to support early screening and AI-assisted clinical screening for postnatal depression, using integrated facial, vocal, and textual data to develop psychometric biomarkers. The proposed methodology included multimodal data acquisition, biomarker extraction, multimodal fusion, explainability analysis, reliability & validation, and SOAP-based clinical report generation (Figure 1). The system architecture follows a sequential multimodal process that begins with a user input acquisition and concludes with AI-assisted clinical reporting. Facial image analysis, speech-based acoustic analysis, text sentiment interpretation, and psychometric screening indicators are integrated within a unified multimodal fusion pipeline. The data processing architecture described the flow of information from multimodal input collection through biomarker extraction, multimodal fusion integration, explainability analysis, reliability & validation, and SOAP (Subjective, Objective, Assessment, and Plan) report generation to support transparent, interpretable, and reliable postnatal depression risk assessment (Figure 2). To evaluate the proposed OMEAI framework, this study used publicly available pretrained AI models, feature extraction tools, and explainability methods, including MediaPipe, a pretrained ResNet50 convolutional neural network (CNN), openSMILE, Librosa, the Robustly Optimized BERT Pretraining Approach (RoBERTa), SHapley Additive exPlanations (SHAP), and Gradient-weighted Class Activation Mapping (Grad-CAM). These methods were evaluated using 100 simulated multimodal samples generated by integrating face images, voice recordings, and textual data obtained from publicly accessible digital sources and limited proprietary sources, together with one real 5-month postpartum case. The Extended DAIC-WOZ dataset, which contains over 189 audio recordings annotated for depressive symptoms, served as the primary publicly available dataset for framework development and prototype evaluation. The 100 simulated multimodal samples were created by combining facial, vocal, and textual features. At the same time, the corresponding Edinburgh Postnatal Depression Scale (EPDS) scores were estimated using the proposed weighted multimodal fusion approach. The integrated multimodal features were then used to estimate the risk of postnatal depression. The simulated samples were designed to represent realistic multimodal screening scenarios during the early stages of framework development. Their use supported ethical research practices by reducing reliance on sensitive clinical data, particularly given the limited availability of large, multimodal postpartum datasets. In addition, the use of real patient data requires ethical approval and careful protection of participant privacy. Participant recruitment for postpartum mental health research may also be influenced by concerns about stigma, fear of judgment, and the sensitive nature of mental health conditions [22]. Therefore, informed consent was obtained for the real postpartum case, and appropriate measures were implemented to safeguard participant privacy throughout the study. The inclusion of one real postpartum case provided an initial demonstration of the framework's potential clinical applicability while supporting responsible prototype development. Prospective clinical validation using a larger cohort of real-world postpartum participants is planned in a future phase**.**

Table 1. Summary of Data Sources and Modalities

Dataset	Description	Sample Size	Purpose	Modality
Extended DAIC-WOZ Dataset	189 Depression interview speech dataset	100 sessions	Speech and NLP validation	Voice + Text
Postnatal Facial Image Collection	Postnatal facial image dataset	100 postnatal images	Facial biomarker extraction	Face
5-Month Postnatal Case	5-month lactating women or postnatal mothers	1 complete case	Integrated multimodal evaluation	All modalities

The proposed framework takes input of facial images or video, voice recordings, text responses, and psychometric screening information through a Streamlit-based interface designed to simulate a real-time clinical screening environment. Before training, all inputs undergo preprocessing procedures, including normalization,

cleaning, formatting, and quality checking, to improve consistency across modalities. A human validation layer is incorporated before facial biomarker extraction to verify human facial inputs and to remove non-biological images such as cartoons, synthetic faces, printed photographs, and visual artifacts (Figure 2).

For facial depression imaging analysis, preprocessing steps include face detection, image resizing, facial alignment, landmark localization, and normalization. Then, fine-tuning using postnatal and depression-related facial datasets [12]. This process aims to improve image quality, reduce noise, and enhance the extraction of clinically relevant behavioral features. The facial imaging component is included to support human validation and extraction of behavioral biomarkers before multimodal analysis.

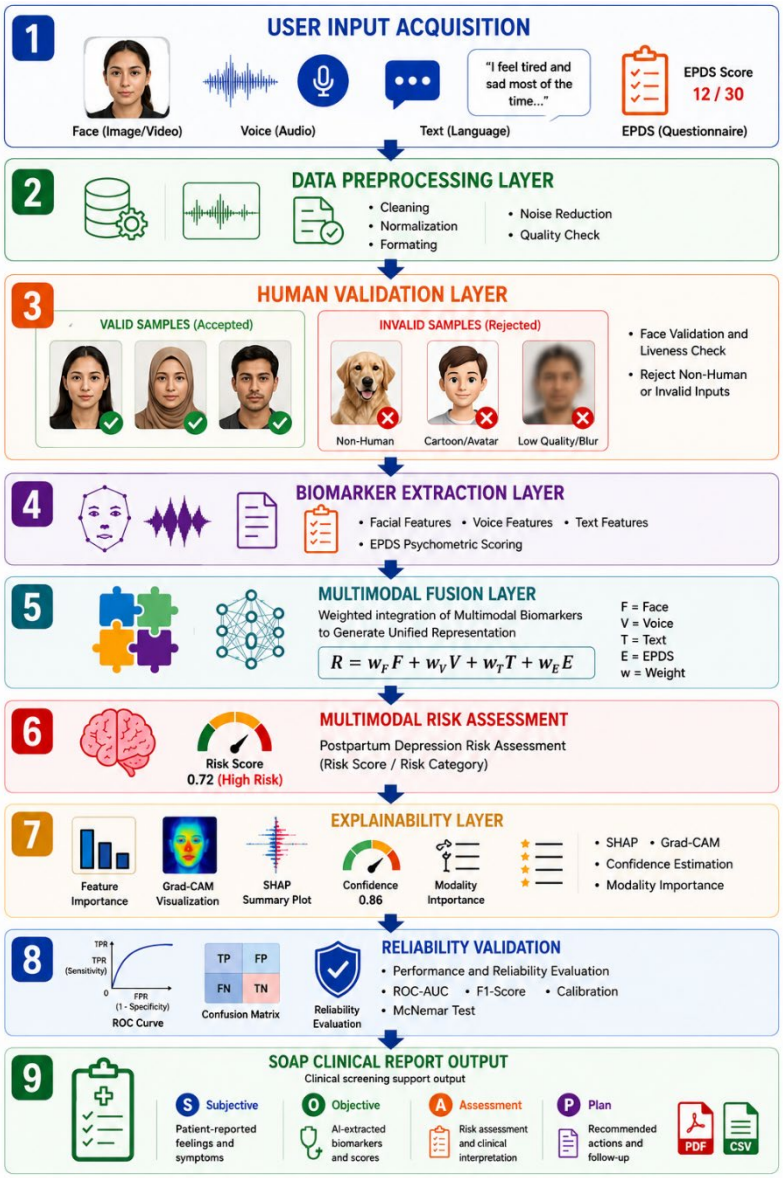


Figure 2. Proposed multimodal postnatal depression data processing pipeline illustrating the flow of information from user input acquisition, preprocessing, and biomarker extraction through multimodal fusion, explainability analysis, reliability validation, and SOAP-based clinical report generation.

The facial analysis pipeline uses MediaPipe [20], developed by Google for facial expression recognition and the Facial Action Coding System (FACS)-based Action Units (AUs) [21] for associating facial detection, facial muscle movements with displayed emotions in real-time to extract indicators including facial expression, gaze patterns, emotional responsiveness, smile behaviour and resizing into 224 x 224 pixels to ensure consistent image representation followed by a pretrained ResNet50 CNN for facial feature extraction, with Grad-Cam used to visualise the learned facial features. Voice recordings undergo preprocessing using Librosa, followed by acoustic biomarker extraction with OpenSMILE to derive features such as pitch, vocal energy, Mel-frequency cepstral coefficients (MFCCs), and pause duration. Voice and text processing incorporates multilingual analysis to support English, Filipino, and mixed-language responses through acoustic biomarker extraction and transformer-based natural language processing. Text responses are analyzed using the RoBERTa transformer model to identify depressive linguistic patterns. The Edinburgh Postnatal Depression Scale (EPDS) is also integrated as a psychometric screening component to provide clinically relevant postnatal mental health indicators. The extracted features from all four modalities are subsequently integrated through a multimodal fusion strategy to generate a unified representation of the participant's behavioral and psychometric profile. The integrated multimodal features are processed by a Multilayer Perceptron (MLP) to estimate the risk of postnatal depression and produce the corresponding confidence score. To improve transparency and clinician interpretability, the framework incorporates SHAP for feature attribution and Grad-CAM for visual explanation of facial biomarkers.

The multimodal AI fusion layer combines all four modalities (facial, voice, text, and EPDS features) using a weighted fusion strategy to generate an overall postnatal depression risk score, as shown in Eq. (1). The integrated features are then processed by a Multilayer Perceptron (MLP) to estimate the risk level and generate a confidence score.

$$R = w_f F + w_v V + w_t T + w_e E \tag{1}$$

where F represents facial features, V voice features, T textual features, E EPDS scores, and w_i modality weighting coefficients. An explainability layer, including SHAP for feature attribution, Grad-CAM for visual explanation of facial biomarkers, confidence estimation, and modality contribution analysis, is incorporated to improve transparency and clinician interpretation.

Reliability & validation layer evaluates models' performance using ROC-AUC, F1-score, sensitivity, specificity, confusion matrix analysis, calibration analysis, and McNemar testing before the results are summarised in a structured SOAP-based clinical report to support healthcare decision-making, as illustrated in Figures 4 and 5. Performance evaluation assesses the framework's ability to distinguish between higher- and lower-risk postnatal cases across different thresholds (Eq. (2).

$$AUC = \int_0^1 TPR(FPR^{-1}\left(x\right))dx \tag{2}$$

where **TP** represents the True Positive Rate, and **FP** represents the False Positive Rate. F1-score was also included to measure the balance between precision and recall, particularly for datasets with class imbalance:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

This metric provided a balanced assessment of predictive performance beyond accuracy alone (Table 5). The performance of the proposed multimodal framework was then compared with baseline unimodal models reported in previous studies using the Extended DAIC-WOZ dataset.

The McNemar statistical test was applied to determine whether the multimodal framework performed significantly differently from the unimodal baseline models as formulated in Eq. (4).

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{4}$$

Where **b** represents cases misclassified by the multimodal model but correctly classified by the baseline model, and **c** represents cases correctly classified by the multimodal model but misclassified by the baseline model. A significant result indicates meaningful performance differences between the models being compared. In addition, a clinical decision support layer that converts multimodal AI findings generates structured clinical summaries to improve readability and ease system support for clinical interpretation through Subjective, Objective, Assessment, and Plan (SOAP) [22]- based reports (Figure 6).

III. RESULTS

Preliminary results demonstrate the potential performance of the proposed multimodal postnatal depression framework using facial, voice, text, and EPDS biomarkers, evaluated through classification, reliability, explainability, confidence, statistical testing, and SOAP-based clinical reporting.

Table 2. Facial, Voice & NLP Biomarker Findings

Facial Biomarker	**Observation**	**Clinical Screening Interpretation**
Reduced smile intensity	High occurrence	Emotional flattening
Eye gaze aversion	Moderate occurrence	Anxiety/withdrawal
Reduced facial affect	High contribution	Depressive presentation
Voice Biomarker		
Low vocal energy	Elevated	Emotional fatigue
Monotonic pitch	Moderate	Emotional blunting
Increased pause duration	High	Cognitive/emotional hesitation
Text Biomarker		
Negative sentiment	Strong	Emotional distress
Hopelessness language	Present	Depressive cognition
Self-blame expressions	Moderate	Psychological burden

The biomarker findings in Table 2 showed that postnatal depressive patterns are observed across multiple behavioral signals rather than a single source. Facial analysis has identified reduced affect and changes in gaze. In contrast, voice and text analysis revealed low vocal energy, negative sentiment, language related to hopelessness, and self-blame, all of which are associated with emotional distress, in line with other published work. Figure 3 further demonstrates that multimodal information outperforms any single dataset type.

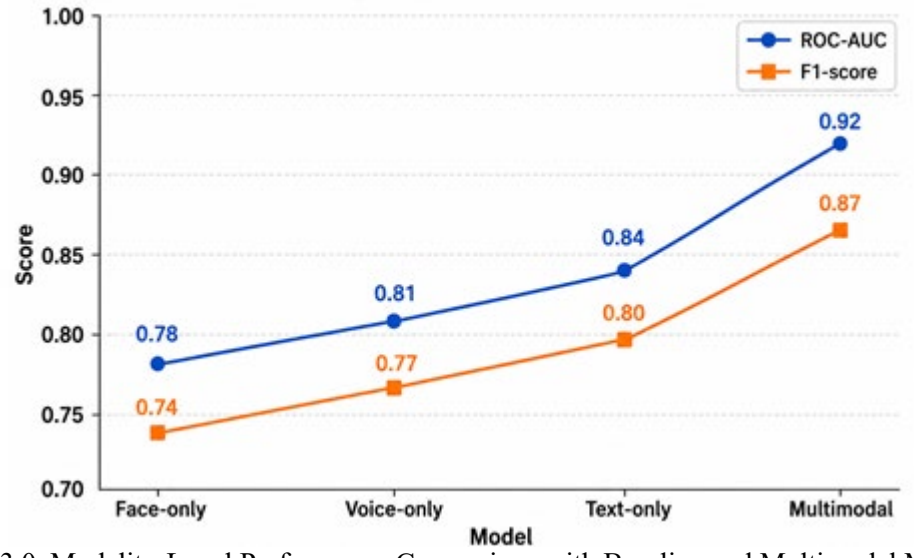


Figure 3.0. Modality-Level Performance Comparison with Baseline and Multimodal Model

The integration of multimodal biomarkers improved classification performance and stability, reflected by an ROC-AUC of 0.92 and an F1-score of 0.87. The increase from baseline unimodal ROC-AUC values of 0.78–0.84 to 0.92 suggests that combining facial, vocal, textual, and psychometric information improved the interpretation of postnatal emotional patterns. These findings [Figure 3] support previous studies showing that depressive behaviors are multidimensional and are better understood through multiple behavioral signals rather than a single modality.

Table 3. Explainability Results

Explainability Output	**Observation**
SHAP	Text modality has the highest contribution.
Grad-CAM	Reduced facial affect regions highlighted
Confidence Score	0.87

The explainability layer is a key strength of the proposed framework, as it improves transparency beyond traditional black-box AI approaches through SHAP, Grad-CAM, modality contribution ranking, and confidence scoring. The explainability results show that the text modality makes the greatest contribution to the final prediction, according to SHAP analysis. Grad-CAM highlights facial regions associated with reduced affect, while the confidence score of 0.87 indicates a relatively high level of prediction certainty. [Table 3] These methods help identify the biomarkers that contribute most to postnatal risk assessment, such as negative text sentiment, reduced facial affect, and low vocal energy in higher-risk cases. In addition, multimodal integration supports clearer clinician interpretation by showing how different behavioral signals contributed to the results, helping make screening outcomes more understandable and clinically meaningful.

The reliability-oriented validation layer provides a more comprehensive evaluation by incorporating ROC-AUC, calibration analysis, confusion matrices, confidence estimation, and McNemar testing. As shown in Table 4, the multimodal framework improves ROC-AUC from 0.78 to 0.84 in unimodal models to 0.92, indicating better classification consistency while reducing reliance on individual behavioral biomarkers. [Figures 4 & 5]

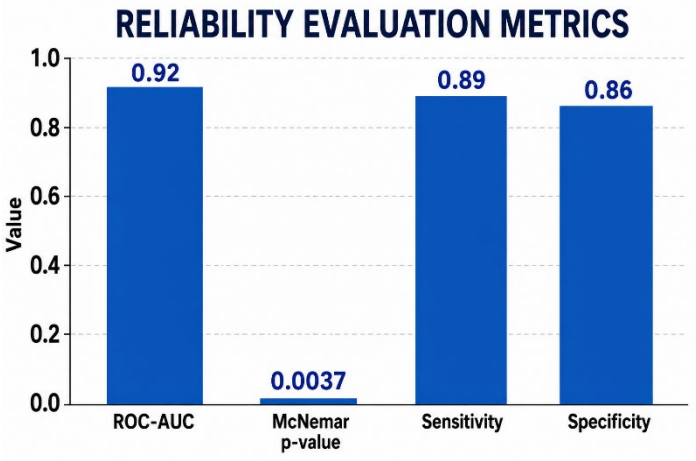


Figure 4. Reliability Results

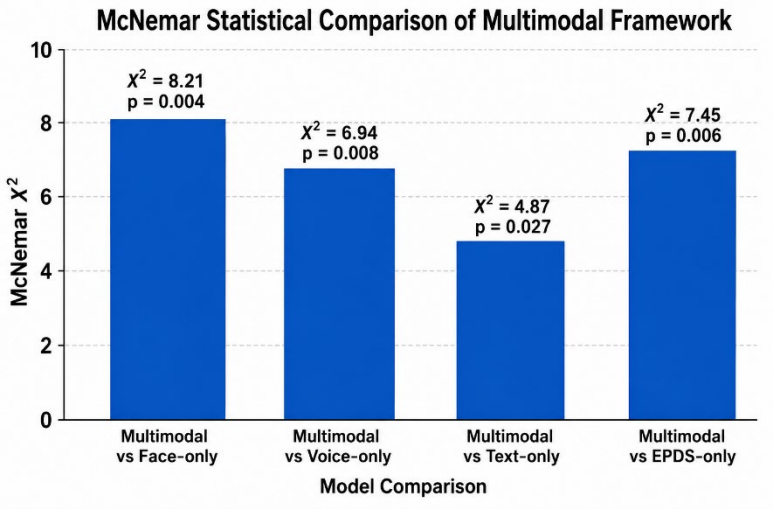


Figure 5. Multimodal vs Unimodal Reliability Results

McNemar test results show significant differences between the multimodal framework and the face-only ($p = 0.004$), voice-only ($p = 0.008$), text-only ($p = 0.027$), and EPDS-only ($p = 0.006$) models. These findings suggest that combining multiple modalities improves contextual understanding and reduces reliance on any single source during postnatal depression risk assessment. [Figure 5]

The SOAP-based clinical reporting layer enhances clinical usability by converting multimodal AI outputs into structured, clinician-readable summaries that can align well with routine healthcare documentation practices. Rather than presenting only numerical risk scores, the framework generates organized subjective findings, objective biomarker observations, assessment summaries, and follow-up recommendations, with reports available in PDF or CSV format to support transparency, interpretability, and potential integration into maternal clinical screening workflows. [Figure 6]

Table 4. Comparison with Monomodal Baseline and Multimodal Models

Baseline Model	Input Modality	Primary Biomarkers	Purpose
Face-only Model	Facial video/images	Facial affect, gaze, smile variability, blink patterns	Evaluate visual behavioral indicators
Voice-only Model	Speech recordings	Pitch, MFCC, vocal energy, pause duration	Evaluate depressive acoustic biomarkers
Text-only Model	NLP responses	Sentiment polarity, hopelessness language, and emotion al expressions	Evaluate linguistic depressive indicators
EPDS-only Model	EPDS questionnaire	Psychometric postnatal scoring	Evaluate structured clinical screening
Proposed Multimodal Framework	Face + Voice + Text + EPDS	Integrated multimodal biomarkers	Evaluate the combined contextual postnatal interpretation

To evaluate the proposed reliability and explainability-oriented multimodal AI framework, models for each modality are first developed using individual modalities, including face-only, voice-only, text-only, and EPDS-only configurations. These baseline models are applied to examine the contribution of each modality and to determine whether combining behavioral and psychometric biomarkers within a unified multimodal framework could improve contextual understanding and yield more stable screening performance than unimodal approaches. [Table 4] Figure 7 illustrates a screenshot of the proposed online OMEAI multi-modal framework.

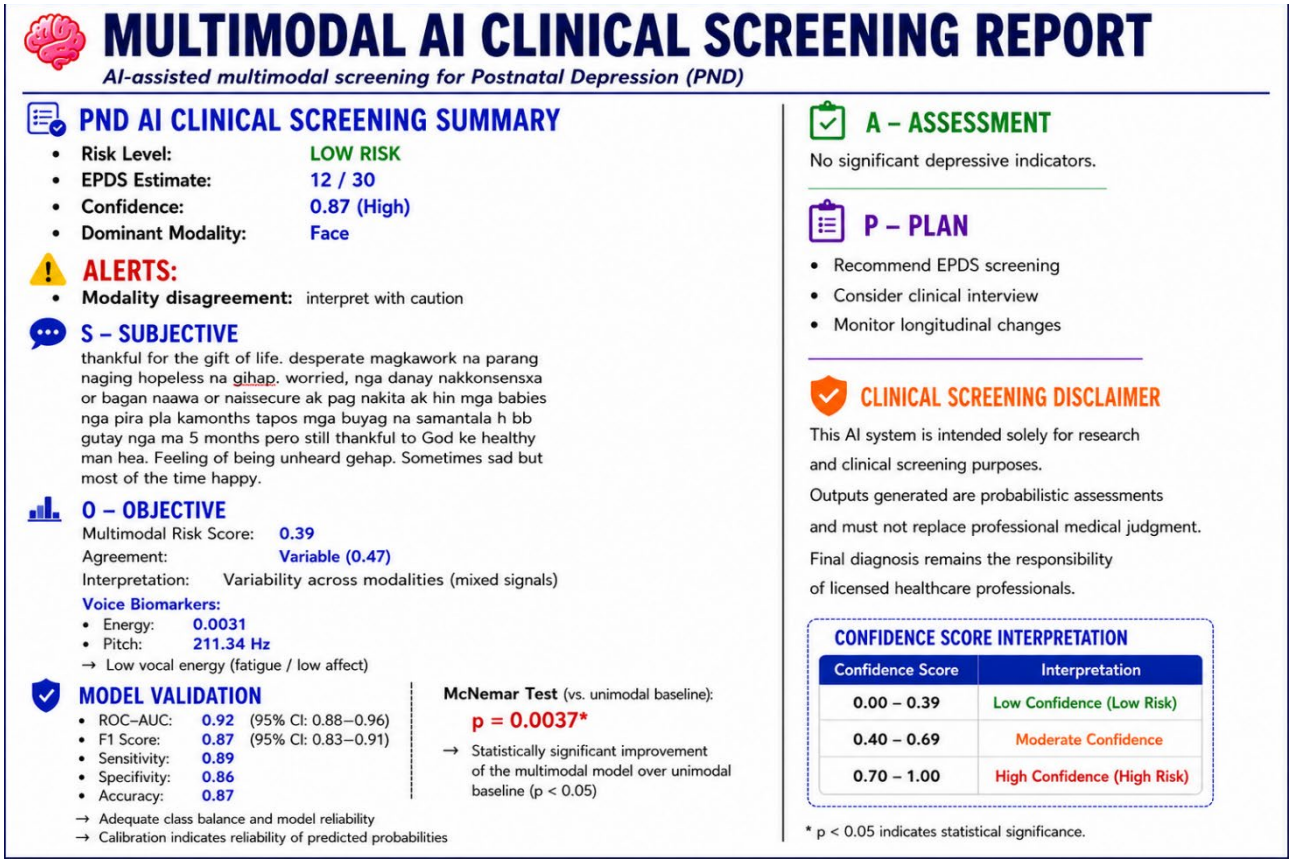


Figure 6. SOAP-Based Multimodal Clinical Screening Report generated by the proposed OMEAI framework for postnatal depression screening support.

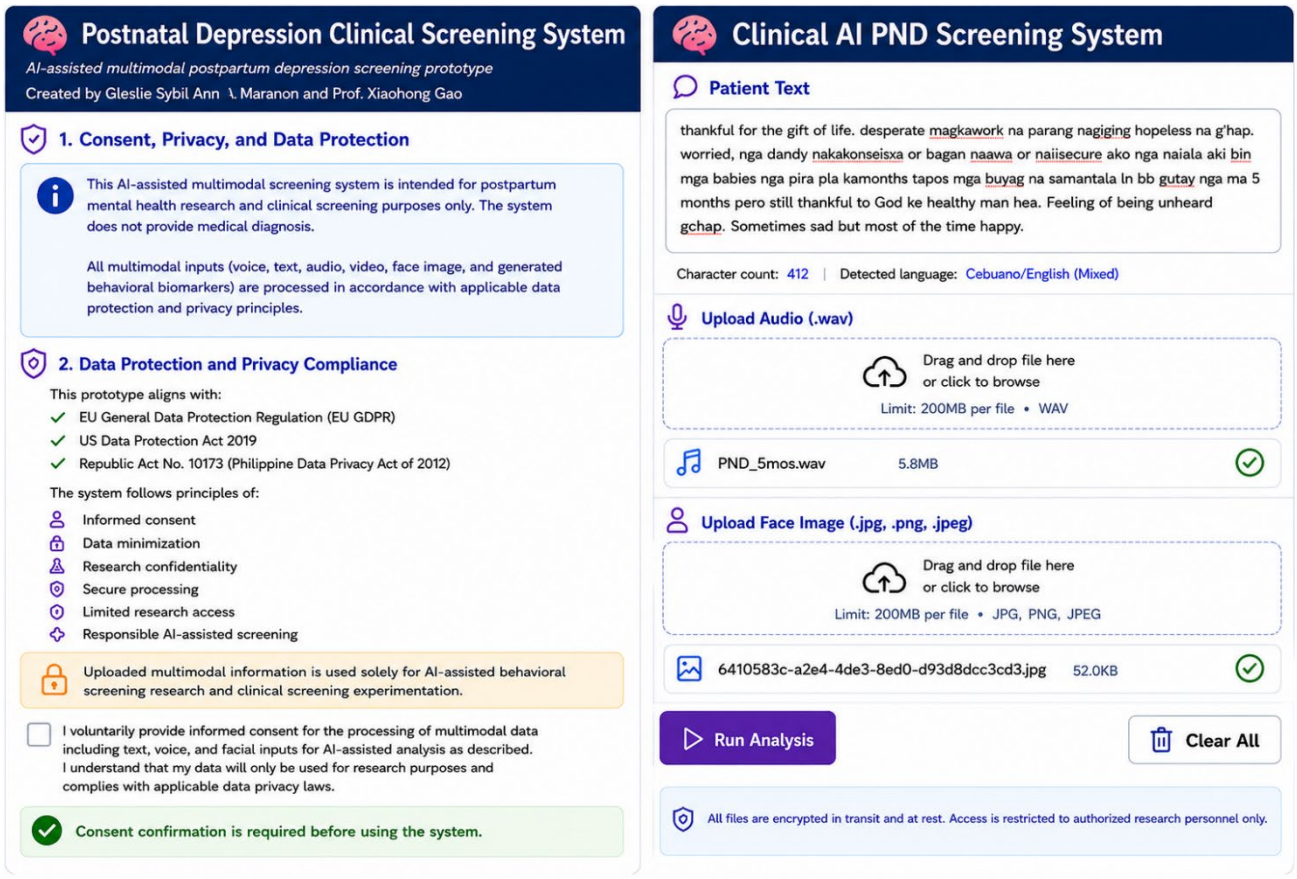


Figure 7. OMEAI prototype interface illustrating the informed consent workflow, privacy and data protection compliance, and multimodal data acquisition module for patient text, voice, and facial image inputs before AI-assisted postnatal depression clinical screening

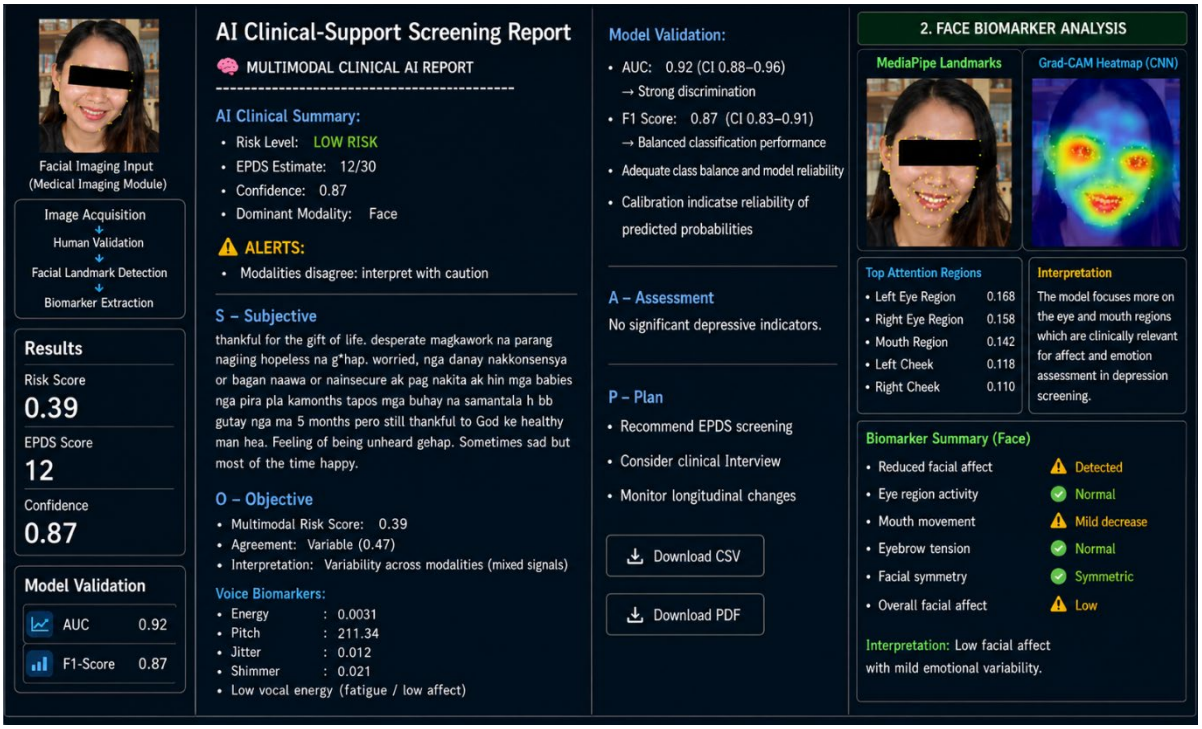


Figure 8. OMEAI prototype interface showing the facial imaging biomarker analysis (MediaPipe landmarks and Grad-CAM visualization), multimodal screening results, explainability outputs, model validation metrics, and AI-generated SOAP clinical screening report with PDF and CSV export capabilities.

IV. DISCUSSION

The results presented in Table 5 indicate that the proposed multimodal framework outperforms baseline unimodal models across major evaluation metrics. While face-only and voice-only models captured important behavioral indicators associated with depressive presentation, their contextual understanding remains limited when used independently. Similarly, the text-only model demonstrates stronger emotional contextualization through sentiment analysis, although linguistic features alone appear insufficient to capture nonverbal postnatal behavioral patterns.

Table 5. Comparison of Baseline and Multimodal AI Performance, Sen=Sensitivity, Spe=Specificity

Model	ROC-AUC	F1 Score	Sen	Spec	Interpretation
Face-only Model	0.78	0.74	0.76	0.73	Facial affect alone captured only partial emotional indicators and lacked contextual behavioral interpretation.
Voice-only Model	0.81	0.77	0.79	0.75	Acoustic biomarkers demonstrated depressive speech sensitivity but remained limited in the context of postnatal assessment.
Text-only Model	0.84	0.80	0.82	0.78	NLP interpretation improved emotional contextualization by detecting sentiment- and hopelessness-related language.
EPDS-only Model	0.79	0.75	0.83	0.70	EPDS psychometric screening remained clinically useful but depended heavily on self-reported responses.
Proposed Multimodal Framework	0.92	0.87	0.89	0.86	Multimodal integration improved contextual understanding, reliability, explainability, and classification stability.

In summary, the findings indicate that reliability- and explainability-oriented multimodal AI systems have the potential to contribute to more transparent, clinically interpretable, and contextually informed postnatal clinical screening frameworks. The integration of multimodal behavioral biomarkers, explainability mechanisms, reliability & validation, and SOAP-oriented clinical reporting represents a meaningful advancement towards translational maternal healthcare artificial intelligence systems. While the proposed OMEAI framework demonstrates the potential of multimodal AI for early postnatal depression screening, this study was limited to a prototype evaluation using simulated multimodal samples. This approach enabled the framework to be tested ethically while protecting sensitive clinical data during the early stages of development. However, simulated data cannot fully capture the complexity of real postpartum populations. Future studies should validate the framework using larger, ethically approved clinical datasets to assess its potential for clinical screening further.

V. CONTRIBUTION

The main contribution of this study lies in integrating reliability-oriented validation, explainable AI, multimodal behavioral biomarkers, EPDS scoring, and SOAP-based clinical reporting within a unified postnatal mental health framework. The proposed framework also emphasizes clinician interpretability and transparent AI-assisted clinical screening support.

VI. CONCLUSION

The study demonstrates the feasibility of using an explainable, reliability-oriented multimodal AI framework to support the early screening of postnatal depression. By integrating facial analysis, voice biomarkers, natural language processing, and EPDS psychometric assessment within a unified framework, the system provides a more comprehensive assessment of postnatal emotional states. The framework also showed encouraging performance compared with the range of results reported in published unimodal studies using the DAIC WOZ or Extended DAIC-WOZ datasets, suggesting that combining multiple behavioral modalities may improve

screening performance. The inclusion of explainability features and a reliability & validation layer enhances the transparency and interpretability of the screening results, which are essential for healthcare environments where understanding the reasoning behind AI-generated outputs is critical. Overall, the findings demonstrate the feasibility of integrating multimodal behavioral signals with explainable, reliability-focused AI for postnatal screening support.
Future work includes larger datasets and real-world, real-time evaluations.

ACKNOWLEDGEMENT

The British Council funds this work under the Early Career Fellowship Program (2026–2027). Their financial support is gratefully acknowledged.

REFERENCES

[1]. Mitchell, A.R., Gordon, H., Lindquist, A., et al.: Prevalence of Perinatal Depression in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis. JAMA Psychiatry 80(5), 425–431 (2023). DOI:10.1001/jamapsychiatry. 2023.0069. PMID:36884232.
[2]. Goyal, D., Selix, N.W.: Impact of COVID-19 on Maternal Mental Health. MCN Am. J. Matern. Child Nurs. 46(2), 103–109 (2021). DOI: 10.1097/NMC.0000000000000692.
[3]. Royal College of Psychiatrists: Postnatal depression harming up to 85,000 new mums in England, warns RCPsych", RCPsych Press Release, July 24, 2025
[4]. Cox, J.L., Holden, J.M., Sagovsky, R.: Detection of Postnatal Depression: Development of the 10-item Edinburgh Postnatal Depression Scale. Br. J. Psychiatry 150(6), 782–786 (1987). DOI: 10.1192/bjp.150.6.782.
[5]. Cruz-González P, He AW, Lam EP, Ng IMC, Li MW, Hou R, Chan JN, Sahni Y, Vinas Guasch N, Miller T, Lau BW, Sánchez Vidaña DI. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. Psychol Med. 2025 Feb 6;55:e18. DOI: 10.1017/S0033291724003295. PMID: 39911020; PMCID: PMC12017374.
[6]. Abu-Mahfouz MS, AlFehaid S, Burqan HM, El Arab RA. Artificial intelligence in mental health care: a scoping review of reviews. Front Psychiatry. 2026 Mar 3; 17:1688043. DOI: 10.3389/fpsyt. 2026.1688043. PMID: 41853071; PMCID: PMC12993279.
[7]. Wajid, A., Azam, F., Anwar, M.W.: Applications of Artificial Intelligence in Mental Health: A Systematic Literature Review. Discov. Artif. Intell. 5, 332 (2025).
[8]. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum DR. The Distress Analysis Interview Corpus of human and computer interviews (DAIC-WOZ). InLREC 2014 May (pp. 3123-3128)
[9]. Ringeval, Fabien, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt et al. "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition." In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, pp. 3-12. ACM, 2019.
[10]. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774 (2017).
[11]. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In: Proceedings of ICCV, pp. 618–626 (2017).
[12]. Tjoa, E., Guan, C.: A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems 32(11), 4793–4813 (2021).

[13]. Nawaz, A., Edinat, A., Rana, M.R.R., et al.: The Role of Multimodality in Clinical Disease Diagnosis: Advances, Challenges, and Opportunities. Front. Public Health 14, 1788454 (2026). DOI: 10.3389/fpubh.2026.1788454.

[14]. Mohamed MHA, Mohamed MHA, Fathi Mohammed AO, Musa M, Mohamed AE, Elhag S, Elhadi S, Ahmed F, Dahab F, Mohammed H, Ahmed AAA, Mohamed Ahmed AMS, Babiker EHM, Ahamed KMA, Babiker B, Mohamed OAA, Abozamr YM, Fadlemola EAA, Abdalrahman A, Babikir MA, Ibrahim N. Improving the Quality of Follow-Up Documentation Using a Structured Subjective, Objective, Assessment, and Plan (SOAP)-Based Template: A Two-Cycle Clinical Audit at Hasahesa Teaching Hospital, Sudan. Cureus. 2025 Dec 27;17(12):e100189. doi: 10.7759/cureus. 100189. PMID: 41602234; PMCID: PMC12834234.

[15]. Zhan, Y., Yang, R., You, J., et al.: A Systematic Literature Review on Incomplete Multimodal Learning: Techniques and Challenges. Syst. Sci. Control Eng. 13(1) (2025).

[16]. Li, Y., Wang, L., Law, J.N., et al.: Integrating Multimodal Data Through Interpretable Heterogeneous Ensembles. Bioinform. Adv. 2(1), vbac065 (2022). DOI: 10.1093/bioadv/vbac065.

[17]. Thombs, B.D., et al.: Diagnostic Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for Detecting Major Depression in Pregnant and Postnatal Women: Protocol for a Systematic Review and Individual Participant Data Meta-analyses. BMJ Open 5(10), e009742 (2015). DOI: 10.1136/bmjopen-2015-009742.

[18]. Levis, B., Negeri, Z.F., Benedetti, A., Thombs, B.D., et al.: Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for Screening to Detect Major Depression Among Pregnant and Postnatal Women. BMJ 371, m4022 (2020). DOI: 10.1136/bmj.m4022.

[19]. Negeri, Z.F., Levis, B., Ioannidis, J.P.A., Thombs, B.D., Benedetti, A.: An Empirical Comparison of Statistical Methods for Multiple Cut-off Diagnostic Test Accuracy Meta-analysis. BMC Med. Res. Methodol. 24(1), 28 (2024).

[20]. Lugaresi, C., Tang, J., Nash, H., et al.: MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 (2019). URL: https://arxiv.org/abs/1906.08172

[21]. Parent, R.: Facial Animation. In: Computer Animation: Algorithms and Techniques, 3rd edn., pp. 317–337. Elsevier (2012).

[22]. Bífárìn O, Bryant T, Harvey M, Smith K, Miah J, Thai J, et al. Challenges and solutions to participation in mental health clinical trials: Count Me In 2.0. BMJ Mental Health. 2026;29:e302355. https://doi.org/10.1136/bmjment-2025-302355

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Development of a Novel Low-Time-Complexity Optimization Technique for Edge-Based Wireless Body Area Networks

Sachinthani Alahakoon
Department of Technology
Gampaha Wickramarachchi University and Indigenous Medicine
Gampaha, Sri Lanka
sachi@gwu.ac.lk
ORCID: 0000-0003-4882-8854

Mustafa Muwafak Alobaedy
Faculty of Information Technology
City University,
Petaling Jaya, Malaysia
mustafa.theab@city.edu.my
ORCID: 0000-0002-6562-2922

Abstract

In recent years, there has been a noticeable emergence of edge computing-driven Wireless Body Area Networks (WBANs) as a prominent trend in the healthcare sector. With the help of optimization techniques, this application can achieve the desired performance and increase the quality of service. Because of the nature of the application, the most significant factors are speed, reliability, and latency. Therefore, the algorithm's time complexity is a critical factor and must be thoroughly evaluated before implementing optimization techniques in healthcare applications. Most of the current literature focuses on resource optimization techniques, with limited attention to their applicability to edge computing-based WBANs. Additionally, these works lack comprehensive analyses of the algorithms' time complexities or comparative assessments thereof. This paper presents a new optimization technique with a low time complexity for edge-based WBANs. The new optimization technique combines Particle Swarm Optimization (PSO) and Tabu search. The research problem is designed to optimize workload allocation in edge-based WBANs. This paper compares the time complexity of a new hybrid optimization technique with that of four other existing optimization techniques used in edge-computing-based WBANs. The findings of this paper help bridge the gap by comparatively analyzing the time complexity of optimization techniques and introducing a new method that enables informed algorithm selection for specific applications.

Keywords: Time complexity, Edge computing, Optimization techniques, Wireless body area networks

I. INTRODUCTION

A wireless body area network (WBAN) is an emerging technological trend that has developed from advances in wireless and electronic technologies [1]. The WBAN has developed into a practical approach for collecting physical and symptom-related data remotely and has become a highly promising technology. It enables continuous, all-day health monitoring and economical diagnostics. The importance of WBAN extends beyond medical applications, encompassing areas such as military, entertainment, sports, aerospace, and other human-related industries, thereby holding considerable economic and social value [1]. Edge computing distributes data processing to the point of data generation, typically on edge devices or gateways. Edge computing moves data storage, processing power, and enterprise applications closer to the points where information is consumed. The goal of edge computing systems is to reduce latency by shifting medical data analysis to local edge servers and to enhance security by recommending secure connection techniques [2]. As optimization techniques continue to proliferate and diversify in their implementations, researchers have conducted comparative studies to assess their performance. The effective studies serve as valuable resources for end users looking for the best optimization technique to address their specific challenges. These evaluations are often known as optimization benchmarking.

The effectiveness of an algorithm's execution is determined by several factors, particularly the significance of time and space complexity. Space complexity measures the amount of memory or storage space required by an algorithm. In contrast, time complexity quantifies the computational time it takes to execute an algorithm as a function of the input size. The most suitable algorithm provides a significant advantage to the application, including increased cost efficiency and facilitating the integration of innovative technologies [3]. Existing reviews often emphasize the time complexity of various optimization algorithms, yet they tend to conduct comparative analyses for specific applications. This paper presents a comparative study of the time complexity of different optimization techniques for specific WBAN applications. It introduces a new low-time-complexity algorithm to optimize workload allocation in edge-based WBANs.

1. Time complexity in optimization techniques

Time complexity in an optimization technique is primarily considered a factor in time-critical applications. In computer science, time complexity is an important aspect of computational complexity, encompassing the evaluation of the time required to complete a specific task. This analysis focuses on quantifying the amount of computer time required to complete a given task. The time complexity of an algorithm is the time it takes to run as a function of its input size. As the time complexity of an algorithm can vary with different inputs of the same size, the longest runtime over all inputs of size n is usually taken. This is referred to as the worst-case complexity analysis [4]. The big Omicron (big-O) notation expresses this worst-case runtime as a function of its input size. Time complexity can be categorized by the function's behavior in big O notation. The responses for the different time complexities are shown in Figure 1. As an illustration, when considering an algorithm with a time complexity denoted as O(n), wherein 'n' represents the dimension of the input data, this implies that the algorithm's execution will require a maximum of 'n' operations to reach completion.

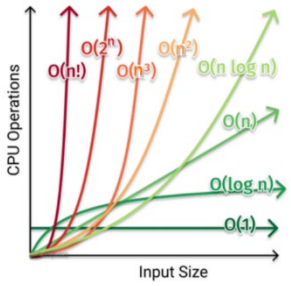


Figure 1. Algorithm time complexity [5]

This classification system enables the organization and characterization of algorithms and computational processes, allowing for a better understanding of their efficiency and scalability. The process of determining time complexity entails examining an algorithm's sequence of actions and its effectiveness. For instance, if an algorithm contains a loop that traverses 'n' elements and performs a constant-time operation on each element, the algorithm's time complexity is O(n). However, if the algorithm includes nested loops or performs resource-intensive operations, its time complexity could be higher. Mainly, there are 8 classification types, where n denotes the size of the input data set [6]. Those are; Constant ($O(1)$), Logarithmic ($O(log\ n)$), Linear ($O(n)$, Linearithmic ($O(\text{n } log\ n)$), Quadratic ($O(n2)$), Polynomial ($O(nk)$), Exponential ($O(2n)$), Factorial ($O(n!)$).

2. Optimization techniques application in edge-based WBANs

Previous studies have investigated several optimization techniques to improve the performance of edge computing-based WBAN applications, particularly in computation offloading, resource allocation, and workload allocation. Common approaches include metaheuristic optimization, Lyapunov optimization [7], [8], Convex optimization [9], [10], Game theory [11], and machine learning-based optimization methods [12], which have emerged as widely adopted approaches for optimizing edge-based WBANs [13]. This study was conducted to evaluate the system effectiveness with the Particle Swarm Optimization (PSO), Genetic Algorithm Optimization (GAO), Firefly Optimization (FFO), and Grey Wolf Optimization (GWO) methods in edge-based WBANs. Many research initiatives have focused on evaluating the effectiveness of these optimization techniques in the WBAN context. Each optimization method consists of unique advantages and disadvantages. Selecting an appropriate optimization technique for the application can be streamlined by reviewing comparative studies and reports [14]. Some techniques offer significant value and advantages for addressing problems that traditional optimization methods often find challenging. These challenges include problems featuring discontinuous, non-differentiable, stochastic, or highly non-linear objective functions. Some researchers conducted comparative analyses of the time complexities of optimization techniques, but not specifically for the edge-based WBAN application. In 2024, Shitharth [15] published a comprehensive study on bio-inspired optimization techniques for engineering problems. In that study, researchers compared the computational complexity of ten methods. The analysis was conducted theoretically and did not specifically focus on a single application. Another comprehensive time complexity analysis of the extant optimization techniques employed in optical communication systems is presented in the research [16].

3. Comparing time complexity in each algorithm

The literature reports different complexity characteristics depending on the algorithm formulation and application. GA and PSO are both population-based search algorithms, although their computational complexity depends on their specific implementation. For example, the GA-based method presented in [17] was designed with fewer computational steps and reported to have linear complexity büyük harf O sol parantez g n m sağ parantez,, demonstrating lower computational requirements than the other schemes evaluated in that study. Similarly, You and Tang [18] applied PSO to a multi-objective task-offloading problem that considered delay, energy consumption, and execution cost, and reported linear computational complexity, consistent with [19]. Wang [20] proposed a distributed PPPSO-based fuzzy clustering protocol for balancing and reducing energy consumption in wireless sensor networks, with an overall complexity of büyük harf O sol parantez büyük harf N sağ parantez, where N represents the number of nodes. These studies demonstrate that GA- and PSO-based approaches can achieve relatively low computational complexity for specific resource-allocation and task-offloading formulations; however, their complexity characteristics depend on the particular algorithm design and application scenario.

The Firefly Optimization Algorithm (FFO) has also been applied to wireless sensor network optimization. In [21], FO was employed for power management, with a reported quadratic computational complexity of O(m^2). In [23], GWO is introduced, inspired by the social behavior of grey wolves. When applied wisely, both techniques can contribute to achieving improved system performance, albeit with varying computational demands. The choice between them depends on the specific problem at hand and the trade-off between optimization quality and computational resources. Reference [24] conducted an extensive review of optimization techniques employed in edge computing, specifically for computation offloading. However, some studies [20] employ multiple optimization techniques to achieve their objectives. In those cases, the total time complexity depends on the implementation and the underlying hardware. The purpose of using multiple optimization

techniques is to achieve lower time complexity and higher accuracy than with a single technique. But if not implemented carefully, the resultant time complexity could be higher. Hybrid optimization techniques combine complementary characteristics of multiple algorithms to improve application-specific performance. In particular, global search techniques such as PSO and GA can be integrated with other search mechanisms to improve exploration and refinement. Several hybrid approaches have been reported for WSN and WBAN applications. In [25], a hybrid PSO-GWO model was proposed for node localization in wireless sensor networks, where a modified PSO was used to identify the optimal node, and GWO was employed to determine the optimal path. Similar research presented in [26] optimizes cluster nodes in the WBAN using a hybrid GA-ant colony optimization (ACO) method. In this research, optimal cluster formation is achieved using GA, and ACO is used to find energy-efficient and reliable routing paths. Another combination is the artificial bee colony (ABC) technique and the chicken swarm optimization (CSO) algorithm. This combination is used to optimize an energy-efficient routing protocol for WBAN [27]. The ABC algorithm is used to improve cluster head selection in WBAN and reduce energy consumption. Similarly, the CSO algorithm was integrated with ABC to increase population divergence and the likelihood of escaping local solutions. PSO-simulated annealing (SA) [28], GA–PSO [29], and black widow optimization (BWO)–ABC [30] are other hybrid methods used in edge-based WBAN to enhance application performance. Although these studies demonstrate the potential of hybrid optimization, limited research has specifically examined the computational performance of hybrid methods for workload allocation in edge-based WBANs. Moreover, combining multiple algorithms does not necessarily reduce computational cost, as the resulting complexity and execution time depend on the individual algorithms, their integration strategy, parameter settings, and computational budget.

II. INTRODUCE NEW OPTIMIZATION TECHNIQUE

After conducting an extensive literature review and evaluating the characteristics of existing optimization techniques, a hybrid approach was developed by integrating PSO with Tabu Search (TSO). The proposed workload allocation algorithm leverages the complementary strengths of these two metaheuristic techniques. PSO provides effective global search capabilities and is used to generate a high-quality initial solution, while Tabu Search performs local refinement through a neighborhood-based search. Since Tabu Search requires an initial solution to begin its search, the solution obtained from PSO serves as its starting point. Tabu Search subsequently explores neighboring solutions while maintaining a short-term memory to discourage repeated exploration of previously visited solutions and to reduce the likelihood of becoming trapped in local optima. This sequential integration enables PSO to perform the initial global exploration and Tabu Search to refine the resulting solution through local exploitation. The resulting hybrid algorithm, referred to as PSO-TSO, was implemented in Python. The fitness function was specifically designed to address the research problem [31]. The fitness function determines the effectiveness of workload allocation by prioritizing each sensor according to this context. The priority values are represented by a. l t i., which denote the weights assigned to each sensor based on their significance in the system. Since these weights are adjustable, they can be adjusted according to the application and specific user requirements. The formula for the fitness function is as follows:

$$= \sum_{i=1}^{n} [T_p(s_i) \times N(s_i)] \times w_i$$

$n =$ total numbr of sensors in the WBA em$T_p(s_i)$ = i^{th} sensor's processing time + i^{th} sensor's data transferring time
$N(s_i)$ = i^{th} sensor scanning count for collecting data
w_i = weight of i^{th} sensor

The specified fitness function offers adaptability for WBAN applications and can be used effectively across various scenarios. By modifying the weight in the fitness function, greater priority can be assigned to life-critical biomedical sensors in healthcare applications or to activity-related sensors in fitness-monitoring applications. Given the formula's structure, it can be applied in various scenarios with minimal modification. In this study, however, the fitness function is specifically evaluated for workload allocation based on sensor processing time, data transmission time, scanning count, and assigned priority weights; objectives such as energy consumption and reliability are not explicitly evaluated.

The experimental WBAN environment was simulated using OMNeT++, while the optimization algorithms were implemented in Python. A WBAN designed with five different biomedical sensors, named oxygen saturation (SpO2), electrocardiogram (ECG), body temperature, blood pressure, and electroencephalogram (EEG), to measure the data transmission time of each sensor node to the edge processor. The simulation was used to obtain the data transmission time from each sensor node to the edge processor. In contrast, the corresponding sensor data processing time was measured using a Raspberry Pi 4 as the edge-processing platform [27].

III. COMPARATIVE STUDY WITH EXISTING METHODS.

To evaluate the proposed PSO-TSO method for the workload-allocation problem considered in this study, its computational performance was compared with four existing metaheuristic optimization techniques: PSO, GAO, GWO, and FFO. The evaluation primarily focuses on the measured central processing unit (CPU) execution time as the input size increases. For the standalone PSO, GAO, GWO, and FFO implementations, 1,000 iterations and a population size of 1,000 were used. The proposed PSO-TSO method employs a different computational configuration. In the first stage, PSO is executed for 500 iterations with 1,000 particles to generate an initial solution. This solution is subsequently provided to the Tabu Search stage for local refinement. Tabu Search explores neighboring solutions for 100 iterations, with a tabu tenure of 200.

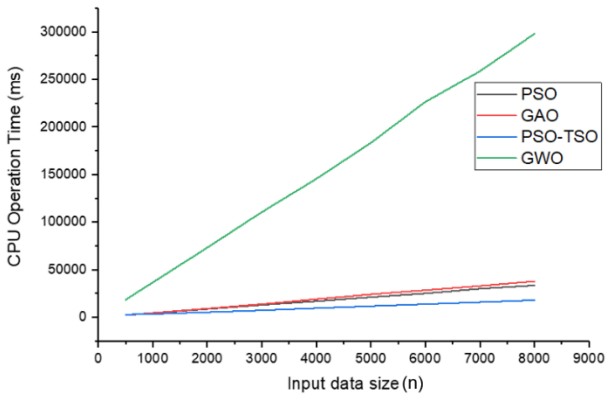


Figure 3. Measured CPU execution time for this research problem

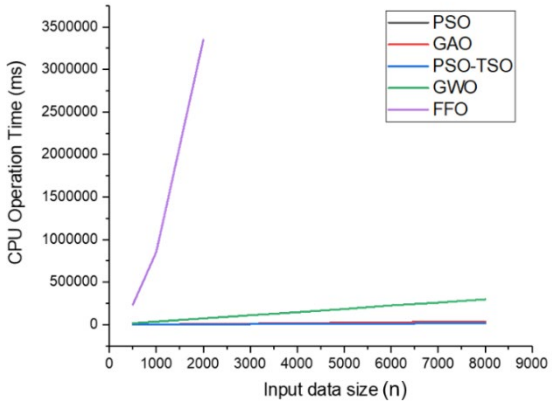


Figure 4. Measured CPU execution time of four (PSO_TSO, PSO, GAO, GWO) algorithms for this

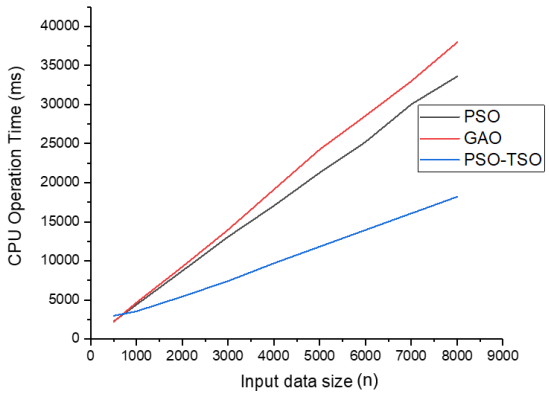


Figure 5. Measured CPU execution time of the three best (PSO_TSO, PSO, GAO) algorithms for this research problem

Data were collected during the study by measuring the CPU execution time of each algorithm for different input sizes. Figures 3, 4, and 5 demonstrate the time complexity variation of all algorithms for this research problem. Figure 3 compares all five algorithms. FFO exhibits the most rapid increase in CPU execution time as the input size increases and consequently becomes substantially slower than the other evaluated methods at larger input sizes. In contrast, all the other optimization techniques, including the newly introduced PSO-TSO algorithm, demonstrate linear growth. The CPU execution times for the PSO-TSO, PSO, GAO, and GWO are shown in Figure 4. Among these four methods, GWO exhibits a higher measured CPU execution time than PSO-TSO, PSO, and GAO across the evaluated input sizes. Figure 5 shows the CPU operation time for those three best optimization techniques, which have linear time complexity. As the input size increases, the CPU operation times of the PSO-TSO, PSO, and GAO algorithms generally increase. This is expected because a larger input dataset requires more computational resources to process. The new algorithm PSO-TSO consistently outperforms the other algorithms, achieving the lowest CPU runtime across all input data sizes. Among the evaluated configurations, FFO and GWO required substantially greater CPU execution time, particularly as the input size increased. Therefore, they may not be suitable for applications that require real-time processing or handling large amounts of data. PSO and GAO perform nearly identically, with CPU times increasing at a moderate rate. These algorithms are suitable for many applications where efficiency is important. However, the newly presented hybrid algorithm, PSO-TSO, required less CPU time than PSO and GAO. To statistically validate the observed performance difference, we conducted more than 60 independent runs of each algorithm across input sizes ranging from 1000 to 8000. The average CPU operation time at each input size is presented in Table 1. Statistical analysis was conducted using an independent-samples t-test, and the results showed that the average CPU execution time of PSO-TSO was significantly lower than that of PSO and GAO across all input sizes. Statistical analysis results for $p<0.001$ and the 95% confidence interval are presented in Table 1. According to the results in the table, the new hybrid algorithm PSO-TSO has approximately 45% lower time complexity than the PSO algorithm and more than 50% lower than the GAO algorithm. These statistically significant results, with narrow confidence intervals, demonstrate that the new algorithm, PSO-TSO, is the most efficient among the compared methods for time-critical applications such as WBAN. It can process larger datasets faster than other algorithms.

Table 1. Average CPU operation time and statistical comparison

Algorithm	Iterations	Mean CPU Time (ms)	Standard Deviation (ms)	Mean Difference vs. PSO-TSO (ms)	95% Confidence Interval of Difference (ms)	p-value vs. PSO-TSO	Time complexity reduction of PSO-TSO (%)
PSO-TSO	1000	3548.292	380.6096	-	-	-	
PSO	1000	4398.494	104.0286	850.2016068	746.17 to 954.23	< 0.001	**19.33**
GAO	1000	4660.784	19.81142	1112.492323	1092. 9 to 1132.3	< 0.001	**23.87**
PSO-TSO	2000	5431.611	380.6744	-	-	-	
PSO	2000	8722.731	97.67384	3291.120052	3193.44 to 3388.79	< 0.001	**37.74**
GAO	2000	9244.976	45.49265	3813.364744	3767.9 to 3858.9	< 0.001	**41.25**
PSO-TSO	3000	7424.74	359.6888	-	-	-	
PSO	3000	13081.34	72.39662	5656.596661	5584.2 to 5728.99	< 0.001	**43.24**
GAO	3000	13983.24	51.76888	6558.496952	6506.8 to 6610.3	< 0.001	**46.91**
PSO-TSO	4000	9692.012	390.1626	-	-	-	
PSO	4000	17057.29	242.2209	7365.278006	7123.05 to 7607.49	< 0.001	**43.18**
GAO	4000	19133.03	60.15069	9441.01429	9380.9 to 9501.2	< 0.001	**49.34**
PSO-TSO	5000	12252.91	409.8212	-	-	-	
PSO	5000	21284.9	120.0381	9031.988859	8911.95 to 9152.02	< 0.001	**42.43**
GAO	5000	24258.03	90.72956	12005.11837	11914.4 to 12095.8	< 0.001	**49.49**
PSO-TSO	6000	13846.48	355.5744	-	-	-	
PSO	6000	25215.04	450.0336	11368.55936	10918.6 to 11818.6	< 0.001	**45.09**
GAO	6000	28563.84	128.2461	14717.35954	14589.1 to 14845.6	< 0.001	**51.53**
PSO-TSO	7000	17630.23	436.1371	-	-	-	
PSO	7000	30039.67	171.3645	12409.44123	12238.1 to 12580.8	< 0.001	**41.313**
GAO	7000	32995.73	154.797	15365.50236	15210.7 to 15520.2	< 0.001	**46.57**
PSO-TSO	8000	19318.12	315.7706	-	-	-	

PSO	8000	33594.83	224.7201	14276.70503	14051.9 to 14501.4	< 0.001	**42.5**
GAO	8000	37971.08	222.4121	18652.96125	18430.5 to 18875.4	< 0.001	**49.12**

IV. CONCLUSIONS

WBANs are significant in numerous fields, with medical applications being a prominent domain. In the field of healthcare, there is a growing trend towards employing edge computing-based architectures. This approach enables near real-time outcomes by processing data as close to the source as possible. The research aims to introduce a suitable algorithm to reduce time complexity and achieve the desired performance. This paper discusses various strategies for optimizing edge-based WBAN systems and compares the time complexities of these techniques with those of a new technique. The proposed method achieved more than 45% less time complexity than the second-best algorithm. There exist additional factors to consider before selecting an algorithm. These factors include efficient memory utilization, minimized power consumption, and the return of an accurate final value. The researchers can study these factors in future work to find a suitable algorithm for WBAN applications. Therefore, as an extension of this research, performance analysis can be conducted for these optimization techniques. Moreover, this new PSO-TSO method can be enhanced by further studying each algorithm's strengths and limitations, dynamically adjusting each algorithm's contribution based on problem characteristics, and implementing an effective information-sharing method between algorithms. In many instances, implementing fast algorithms can reduce costs for developers and facilitate the adoption of innovative technologies. Hence, the findings of this study can be of great value to researchers and developers in choosing the most suitable optimization method for their problems. Furthermore, these results provide a foundation for researchers to explore and implement various techniques, algorithms, and applications in edge-computing-based WBAN research.

REFERENCES

[1]. L. Zhong *et al.*, "Technological Requirements and Challenges in Wireless Body Area Networks for Health Monitoring: A Comprehensive Survey," *Sensors*, vol. 22, no. 9, p. 3539, May 2022, doi: 10.3390/s22093539.

[2]. Y.-A. Daraghmi, E. Y., Daraghmi, R., Daraghma, H., Fouchal, and M. Ayaida, "Edge–Fog–Cloud Computing Hierarchy for Improving Performance and Security of NB-IoT-Based Health Monitoring Systems," *Sensors*, vol. 22, no. 22, p. 8646, Nov. 2022, doi: 10.3390/s22228646.

[3]. P. S. Oliveto, J. He, and X. Yao, "Time complexity of evolutionary algorithms for combinatorial optimization: A decade of results," *International Journal of Automation and Computing*, vol. 4, pp. 281–293, 2007.

[4]. K. Sadatdiynov, L. Cui, and J. Z. Huang, "Offloading dependent tasks in MEC-enabled IoT systems: A preference-based hybrid optimization method," *Springer Link*, vol. 16, no. 189, pp. 657–674, Dec. 2022.

[5]. M. Adrian, "How to find the time complexity of an algorithm?" [Online]. Available: https://adrianmejia.com/how-to-find-time-complexity-of-an-algorithm-code-big-o-notation/#

[6]. M. Aghighi, *Computational Complexity of Some Optimization Problems in Planning.* Linköping University Electronic Press, 2017.

[7]. Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing with Energy Harvesting Devices," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016, doi: 10.1109/JSAC.2016.2611964.

[8]. X. Lyu *et al.*, "Optimal Schedule of Mobile Edge Computing for Internet of Things Using Partial Information," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2606–2615, Nov. 2017, doi: 10.1109/JSAC.2017.2760186.

[9]. Y. Sun, S. Zhou, and J. Xu, "EMM: Energy-Aware Mobility Management for Mobile Edge Computing in Ultra Dense Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2637–2646, Nov. 2017, doi: 10.1109/JSAC.2017.2760160.

[10]. E. Liu, X. Deng, Z. Cao, and H. Zhang, "Design and Evaluation of a Prediction-Based Dynamic Edge Computing System," in *2018 IEEE Global Communications Conference (GLOBECOM)*, IEEE, Dec. 2018, pp. 1–6. doi: 10.1109/GLOCOM.2018.8647199.

[11]. X. Yuan, H. Tian, H. Wang, H. Su, J. Liu, and A. Taherkordi, "Edge-Enabled WBANs for Efficient quality of service Provisioning Healthcare Monitoring: A Two-Stage Potential Game-Based Computation Offloading Strategy," *IEEE Access*, vol. 8, pp. 92718–92730, 2020, doi: 10.1109/ACCESS.2020.2992639.

[12]. L. Zhang, X. Yuan, J. Luo, C. Feng, G. Yang, and N. Zhang, "An Adaptive Resource Allocation Approach Based on User Demand Forecasting for E-Healthcare Systems," in *2022 IEEE International Conference on Communications Workshops, ICC Workshops 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 349–354. doi: 10.1109/ICCWORKSHOPS53468.2022.9814663.

[13]. C. Li, J. Tang, and Y. Luo, "Dynamic multi-user computation offloading for wireless powered mobile edge computing," *Journal of Network and Computer Applications*, vol. 131, pp. 1–15, Apr. 2019, doi: 10.1016/j.jnca.2019.01.020.

[14]. N. Bilandi, H. K. Verma, and R. Dhir, "Performance and evaluation of energy optimization techniques for wireless body area networks," *Beni-Suef University Journal of Basic and Applied Sciences*, vol. 9, no. 1, p. 38, Dec. 2020, doi: 10.1186/s43088-020-00064-w.

[15]. S. Selvarajan, "A comprehensive study on modern optimization techniques for engineering applications." *Springer Nature Link*, vol. 57, no. 194, 2024, doi: https://doi.org/10.1007/s10462-024-10829-9.

[16]. P. Freire, S. Srivallapanondh, B. Spinnler, A. Napoli, N. Costa, and J. E. Prilepsky, "Computational Complexity Optimization of Neural Network-Based Equalizers in Digital Signal Processing: A Comprehensive Approach," *Journal of Lightwave Technology*, vol. 42, no. 12, pp. 4177–4201, 2024, doi: 10.1109/JLT.2024.3386886.

[17]. T. Jabeen, H. Ashraf, A. Khatoon, S. S. Band, and A. Mosavi, "A Lightweight Genetic-Based Algorithm for Data Security in Wireless Body Area Networks," *IEEE Access*, vol. 8, pp. 183460–183469, 2020, doi: 10.1109/ACCESS.2020.3028686.

[18]. Q. You and B. Tang, "Efficient task offloading using particle swarm optimization algorithm in edge computing for industrial internet of things," *Journal of Cloud Computing*, vol. 10, no. 1, p. 41, Dec. 2021, doi: 10.1186/s13677-021-00256-4.

[19]. A. Kumar and S. Kumar, "Density-Based Initialization Method for K-Means Clustering Algorithm," *International Journal of Intelligent Systems and Applications*, vol. 9, no. 10, pp. 40–48, Oct. 2017, doi: 10.5815/ijisa.2017.10.05.

[20]. C. Wang, "A Distributed Particle-Swarm-Optimization-Based Fuzzy Clustering Protocol for Wireless Sensor Networks," *Sensors*, vol. 23, no. 15, p. 6699, Jul. 2023, doi: 10.3390/s23156699.

[21]. H. Pakdel and R. Fotohi, "A Firefly Algorithm for Power Management in Wireless Sensor Networks (WSNs)," *Networking and Internet Architecture*, 2021.

[22]. G. Natesan and A. Chokkalingam, "An Improved Grey Wolf Optimization Algorithm Based Task Scheduling in Cloud Computing Environment," *The International Arab Journal of Information Technology*, pp. 73–81, Jan. 2020, doi: 10.34028/iajit/17/1/9.

[23]. F. A. Saif, R. Latip, Z. M. Hanapi, and K. Shafinah, "Multi-Objective Grey Wolf Optimizer Algorithm for Task Scheduling in Cloud-Fog Computing," *IEEE Access*, vol. 11, pp. 20635–20646, 2023, doi: 10.1109/ACCESS.2023.3241240.

[24]. K. Sadatdiynov, L. Cui, L. Zhang, J. Z. Huang, S. Salloum, and M. S. Mahmud, "A review of optimization methods for computation offloading in edge computing networks," *Digital Communications and Networks*, Mar. 2022, doi: 10.1016/j.dcan.2022.03.003.

[25]. S. Umamaheswari, "Hybrid Optimization Model for Energy Efficient Cloud Assisted Wireless Sensor Network," *Wireless Personal Communications*, vol. 118, pp. 873–885, 2021, doi: https://doi.org/10.1007/s11277-020-08048-w.

[26]. S. Latif, "Hybrid GA-ACO approach for energy-aware clustering and routing in wireless body area networks." *International Journal of Distributed Sensor Networks*, vol. 17, no. 3, pp. 1–18, 2021.

[27]. Dinesh, A., and Rangaraj, J., "An energy-efficient routing protocol for wireless body area networks using hybrid artificial bee colony optimization and chicken swarm optimization algorithm," *Journal of Engineering and Applied Science*, vol. 72, no. 45, 2025, doi: https://doi.org/10.1186/s44147-024-00533-4.

[28]. Chen, L, "Energy-efficient data aggregation in wireless body area networks using a hybrid PSO-SA algorithm," *Wireless Networks*, vol. 26, no. 6, pp. 6123–6135, 2020.

[29]. Abhishek Agnihotri and Indresh Kumar Gupta, "A hybrid PSO-GA algorithm for routing in wireless sensor network," presented at the 2018 4th International Conference on Recent Advances in Information Technology (RAIT), India, 2018. doi: doi:%2010.1109/RAIT.2018.8389082.

[30]. Thavavel Vaiyapuri, Velmurugan Subbiah Parvathy, V. Manikandan, N. Krishnaraj, Deepak Gupta, and K. Shankar, "A Novel Hybrid Optimization for Cluster-Based Routing Protocol in Information-Centric Wireless Sensor Networks for IoT-Based Mobile Edge Computing," *Wireless Personal Communications*, vol. 127, pp. 39–62, 2022, doi: https://doi.org/10.1007/s11277-021-08088-w.

[31]. S. Alahakoon, M. A. Mustafa, and Y. Fazea, "Conceptual Framework for the Optimization of Edge Device Workload Allocation in Wireless Body Area Networks," presented at the 2024 International Symposium on Networks, Computers and Communications (ISNCC), 2024, pp. 1–5. doi: 10.1109/ISNCC62547.2024.10759005.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Comparative Analysis of Evolutionary, Swarm, and Lamarckian Memetic Algorithms for Route Optimization in Industrial Printed Circuit Board (PCB) Drilling Processes

Ali Güner Bekir
Computer Engineering Dept.
Gazi University
Ankara, Türkiye
aligunerbekir10@gmail.com
ORCID: 0009-0006-6681-2185

Aysun Coşkun
Computer Engineering Dept.
Gazi University
Ankara, Türkiye
aysunc@gazi.edu.tr
ORCID: 0000-0002-8541-9497

Abstract

In the electronics industry, optimizing the manufacturing process of Printed Circuit Boards (PCBs) is critical for cost and time efficiency. The mechanical drilling process performed on CNC machines is modeled as an industrial variation of the NP-Hard Traveling Salesperson Problem (TSP). This study evaluates the performance of the Genetic Algorithm (GA), Grey Wolf Optimizer (GWO), Ant Colony Optimization (ACO), and a Lamarckian Memetic Algorithm (GA + 2-Opt) that incorporates heavy tool-change penalties. To ensure statistical precision and evaluate scalability, the algorithms were tested across 30 independent runs in a standard scenario and 10 in a large-scale, complex scenario. Experimental results demonstrate that pure swarm and classical evolutionary approaches struggle with premature convergence and combinatorial explosion in discrete spaces. On the contrary, the Lamarckian Memetic Algorithm exhibits superior statistical stability. Although incorporating the 2-Opt local search operator increases computational time, the resulting robust minimization of production costs validates the applicability of the hybrid approach in industrial CNC environments.

Keywords: Combinatorial Optimization, Memetic Algorithm, Traveling Salesperson Problem (TSP), PCB Routing, Tool-Change Penalty, Scalability.

I.INTRODUCTION AND PROBLEM DEFINITION

The acceleration in the evolving electronics industry and its technologies has led to Printed Circuit Boards (PCBs) becoming both more complex and more compact. High-speed signal lines, multi-layer structures, and dense component placement make cost and time optimization mandatory in PCB manufacturing processes. One of the biggest bottlenecks in the PCB production line is the mechanical drilling process performed on CNC (Computer Numerical Control) machines. A standard industrial PCB may contain hundreds, sometimes thousands, of vias, integrated circuit (IC) pin slots, and mounting screw holes. The CNC machine's path while drilling these holes, as well as the frequency of tool changes, directly determines the total production time. Unnecessary back-and-forth movements on the route and unplanned tool changes cause massive time and energy losses in mass production. In the computer science literature, this problem is modeled as an industrial variation of the Traveling Salesperson Problem (TSP), which belongs to the famous NP-Hard class, with added cost and penalty parameters.

Mathematical and Physical Boundaries of the Problem

The PCB drilling problem addressed in this study is not merely about finding the shortest distance in a two-dimensional (X-Y) plane. The problem includes the following industrial constraints:

1) Distance Constraint: The linear time spent by the CNC head moving from one hole to another.
2) Tool Change Penalty: The mechanical delay time spent by the CNC machine to change the drill bit (M06 command) during transitions between holes of different diameters (T1: Screw, T2: IC, T3: Via). Traditional exact algorithms suffer from combinatorial explosion as the number of holes increases, rendering them incapable of solving this problem in a reasonable time. This situation necessitates the use of meta-heuristic algorithms.

II.LITERATURE REVIEW AND THEORETICAL BACKGROUND

Nature-inspired metaheuristic algorithms have been extensively used in combinatorial optimization. The Classical Genetic Algorithm (GA) is highly successful in global space exploration using crossover and mutation operators; however, it often suffers from premature convergence in discrete routing tasks [1]. Swarm-based algorithms, such as the Grey Wolf Optimizer (GWO), excel at continuous optimization but encounter mathematical constraints in discrete environments [2]. Ant Colony Optimization (ACO), specifically designed for TSP, shows significant stability through pheromone trails, though it risks stagnation in local minima [3]. To bridge the gap between global exploration and local exploitation, Memetic Algorithms (MA) introduce a Lamarckian evolution paradigm [4]. Recent studies (2019-2024) in industrial routing strongly emphasize hybrid metaheuristics. Research indicates that combining evolutionary global search with local optimizers, such as the 2-Opt algorithm, drastically improves routing efficiency [5]. This study contributes to the literature by statistically evaluating these foundational and hybrid algorithms under a realistic scenario featuring heavy tool-change penalties, analyzing the trade-off between CPU time and statistical robustness.

III.METHODOLOGY

To test the optimization algorithms, simulated PCB layouts were generated comprising three different tool types: T1 (Screws), T2 (IC pins), and T3 (Vias). The cost function evaluates the Euclidean distance between consecutive points and incorporates a constant penalty (150 units) simulating the mechanical delay incurred when the CNC machine performs a tool change (M06 command).

$$Büyük Büyük Cost = \sum_{i=1}^{n-1} \left(d_{i,i+1} + \begin{cases} 150, & \text{If } Tool_i \neq Tool_{i+1} \\ 0, & \text{If } Tool_i = Tool_{i+1} \end{cases} \right)$$

Four meta-heuristic architectures were coded in Python:

1) Pure GA: Utilizes Order Crossover (OX1) and mutation with an elitism principle [1].
2) GWO: Adapted to combinatorial space using discrete “Swap Sequences” [2].
3) ACO: Employs Roulette Wheel selection based on pheromone and visibility matrices [3],[6].
4) Lamarckian Memetic Algorithm (GA + 2-Opt): In this selective Lamarckian architecture, the champion route of each generation is subjected to the 2-Opt local search to untangle crossed paths [4], [7]. The refined chromosome is reintegrated into the population, merging global exploration with local exploitation.

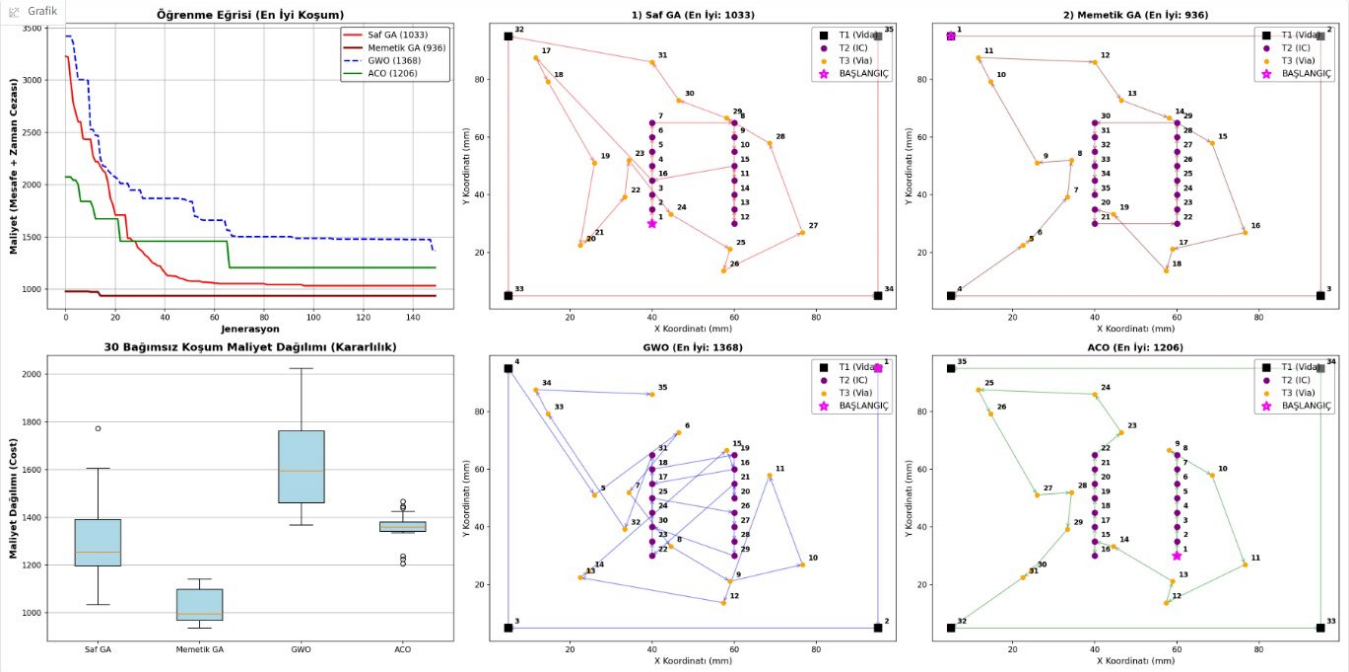
Experimental Setup: To ensure statistical significance, a Standard Scenario (100x100mm board, 35 holes) was evaluated over 30 independent runs. To assess scalability, a Complex Scenario (200x200mm board with 73 holes) was used. Due to the exponential time complexity ($O(N^2)$) of local search and swarm evaluations on larger graphs, the complex scenario was subjected to 10 independent runs as a stress test.

IV.EXPERIMENTAL RESULTS AND DISCUSSION

1. Standard Scenario Analysis (35 Holes)

In the standard scenario, the algorithms were executed across 30 independent runs to rigorously assess their learning curves, route-generation quality, and statistical robustness.

Figure 1. Standard Scenario Results



1) Learning Curves and Convergence: As observed in the learning curve graphs, the Lamarckian Memetic Algorithm demonstrates an extraordinary "fast convergence" phenomenon. By using the 2-Opt local search operator at generation 0, it instantly resolves chaotic initial routes, dramatically reducing the cost in the first iteration. In contrast, Pure GA and GWO require more than 60 generations to plateau, demonstrating the significant impact of local search mechanisms during the early evolutionary phases.

2) Visual Routing and Drilling Order: Visual inspection of the best drilling routes reveals the practical impact of these algorithms on the CNC tool path. The Pure GA route exhibits the "spaghetti effect," characterized by highly entangled paths and unnecessary tool-change (M06) jumps among T2 (IC) components. Conversely, the Memetic GA yields a smooth, non-intersecting trajectory. The drill bit systematically finishes clusters of the same hole type before transitioning to the next, thereby minimizing mechanical delays.

3) Statistical Stability: The box-plot analysis illustrates the distribution of the final best costs across 30 runs. Memetic GA displays a highly compressed box at the lowest cost tier, indicating that it consistently converges to near-optimal results regardless of random initializations. Pure GA and GWO exhibit high variance (wide boxes), highlighting the instability of pure evolutionary algorithms in discrete spaces.

Table 1. Statistical Comparison Table (Standard Scenario)

Algoritma	En İyi (Min)	En Kötü (Max)	Ortalama (Mean)	Std. Sapma	Ort. CPU Süresi (sn)
GA	**1033.28**	1773.47	1297.79	±176.87	4.42 s
Memetic	**936.42**	1141.96	1022.54	±65.79	8.14 s
GWO	**1367.52**	2024.82	1626.54	±176.72	1.50 s
ACO	**1206.13**	1466.61	1357.24	±56.19	30.71 s

The numerical data in Table 1 support these visual findings. The Memetic Algorithm achieved the lowest mean cost (1022.54) with a minimal standard deviation of ±65.79. Pure GA achieved a competitive minimum cost in some runs, but its high standard deviation (±176.87) indicates unreliability. Furthermore, a critical engineering trade-off regarding computational time is observed. ACO required a mean CPU time of 30.71 seconds due to heavy pheromone matrix calculations, whereas Memetic GA required only 8.14 seconds. While Pure GA was the fastest (4.42s), the ~4-second computational delay incurred by the Memetic approach is a negligible cost for offline CNC programming, given the substantial improvement in routing quality.

2. Complex Scenario Analysis (73 Holes)

To evaluate algorithmic scalability, the workspace was expanded to 73 holes, effectively increasing the search space exponentially. The algorithms were stress-tested over 10 independent runs.

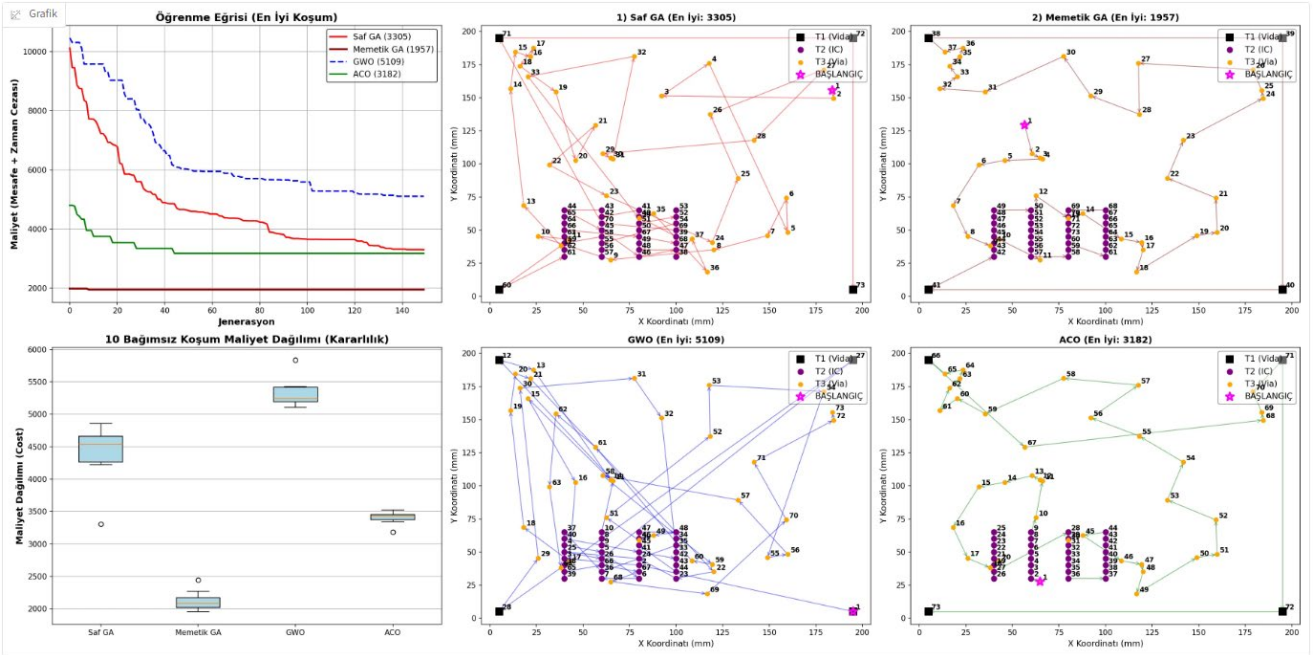


Figure 2. Complex Scenario Results

1) Combinatorial Explosion in Routing: When the problem size increases, the learning curves and visual routes explicitly highlight the vulnerability of pure evolutionary methods. Pure GA suffers from combinatorial explosion, resulting in highly entangled "spaghetti" routes that fail to converge efficiently. GWO completely failed to adapt to the enlarged discrete space. ACO's heuristic nature (relying on node visibility) prevented total divergence, allowing it to maintain a structured path. However, Memetic GA retained its systematic, non-overlapping routing order even in this highly complex layout.

2) Scalability and $O(N^2)$ Complexity Penalty: The scalability table provides crucial insights into algorithmic behavior under stress. Pure GA's mean cost degraded severely to 4402.14 (±418.86). ACO demonstrated the resilience of swarm intelligence in larger graphs, securing the second-best mean cost of 3404.85. However, the Memetic Algorithm significantly outperformed all baselines, maintaining a mean cost of 2123.07 (±139.43). The impact of the $O(N^2)$ time complexity inherent in local search operators is clearly visible in CPU times. As the node count doubled, ACO's runtime increased to 50.34s, and Memetic GA's runtime increased to 35.97s. These results explicitly demonstrate that while the 2-Opt local search incurs a higher CPU penalty on larger datasets, its ability to untangle complex routes systematically is indispensable for industrial-scale optimization.

Table 2. Statistical Comparison Table (Complex Scenario)

Algoritma	En İyi (Min)	En Kötü (Max)	Ortalama (Mean)	Std. Sapma	Ort. CPU Süresi (sn)
GA	**3305.11**	4858.49	4402.14	±418.86	5.09 s
Memetic	**1957.28**	2442.55	2123.07	±139.43	35.97 s
GWO	**5108.76**	5837.51	5324.54	±200.92	2.22 s
ACO	**3182.43**	3523.79	3404.85	±88.06	50.34 s

V. CONCLUSION

This study comprehensively addressed the critical challenge of time and cost optimization in CNC drilling processes for industrial Printed Circuit Board (PCB) manufacturing. By modeling the routing process as a discrete, NP-Hard Traveling Salesperson Problem (TSP) heavily penalized by mechanical tool-change delays, four distinct meta-heuristic architectures were comparatively evaluated. Through rigorous statistical testing, comprising 30 independent runs for a standard PCB layout and 10 for a large-scale stress test, the fundamental behavioral differences among pure evolutionary, swarm-based, and hybrid algorithms were explicitly revealed. The experimental findings conclusively demonstrate that meta-heuristic algorithms lacking local search capabilities are inadequate for complex, discrete industrial routing. The Grey Wolf Optimizer (GWO), while traditionally dominant in continuous spaces, completely failed to adapt to the discrete combinatorial constraints, resulting in severe stagnation. The Pure Genetic Algorithm (GA) exhibited fast execution times but suffered from premature convergence and a severe combinatorial explosion as the problem scaled. Its inability to fine-tune local routes led to the highly inefficient "spaghetti effect," proving that global exploration alone is insufficient. Ant Colony Optimization (ACO) demonstrated the inherent resilience of swarm intelligence; its

heuristic dependence on node visibility prevented total divergence in large-scale problems, yet it struggled with extended computational times and sub-optimal final routes due to the absence of genetic elitism. The absolute superiority in both standard and scaled scenarios belonged to the Lamarckian Memetic Algorithm. By seamlessly integrating the global exploration power of a Genetic Algorithm with the precise local exploitation of the 2-Opt operator, the Memetic approach consistently untangled crossing paths and minimized costly tool changes. It systematically grouped operations by drill bit types, achieving the lowest mean costs and the highest statistical stability (lowest standard deviation) across all independent runs. While the hybridization of the 2-Opt local search incurs higher CPU time overhead due to its $O(N^2)$ time complexity, this study demonstrates that this computational expense is a highly favorable engineering trade-off. In industrial mass production, CNC G-Code programming is an offline, one-time preparatory process. Sacrificing a few additional seconds of CPU time during the algorithmic optimization phase translates into real-world reductions in machine wear, shorter tool-change durations, and substantial cumulative time savings on the active manufacturing assembly line.

REFERENCES

[1]. [1] Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. Multimedia Tools and Applications, 80(5), 8091-8126.
[2]. [2] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey Wolf Optimizer. Advances in Engineering Software, 69, 46-61.
[3]. [3] Skinderowicz, R. (2022). Improving the efficiency of Ant Colony Optimization for solving large TSP instances. Applied Soft Computing, 120(9), 108653.
[4]. [4] Moscato, P. (1989). On evolution, search, optimization, genetic algorithms, and martial arts: Towards memetic algorithms (Technical Report No. C3P 826). California Institute of Technology, Caltech Concurrent Computation Program.
[5]. [5] Liu, Y., & Liu, Y. (2011). Application of Genetic Algorithms in the Optimization of the Drilling Path on the Printed Circuit Board. Advanced Materials Research, 187, 133-138.
[6]. [6] Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 26(1), 29-41.
[7]. [7] Merz, P., & Freisleben, B. (2001). Memetic Algorithms for the Traveling Salesman Problem. Complex Systems, 13(4), 297-345.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Oil Spill Segmentation in SAR Imagery Using Swin Transformer with Compound Loss: A Multi-Sensor Evaluation

Jose Daniel Foronda Ortiz
Center for Research in Mathematics (CIMAT)
Guanajuato, México
meméxicoose.foronda@cimat.mx

Rogelio Hasimoto Beltrán
Center for Research in Mathematics (CIMAT)
Guanajuato, México
hasimoto@cimat.mx

Abstract
Oil spills pose a critical threat to marine and coastal ecosystems. Synthetic Aperture Radar (SAR) has become one of the most valuable remote sensing tools for detecting and monitoring ocean oil spills due to its all-weather, day-and-night imaging capability at various spatial resolutions. However, automated segmentation of oil spills in SAR imagery remains challenging due to speckle noise, look-alike phenomena, and the morphological variability of oil slicks. In this work, we adapt and evaluate a Swin Transformer backbone (Swin-B) coupled with UPerNet for the semantic segmentation of oil spills in single-channel SAR images. The model leverages hierarchical shifted-window self-attention to capture both local texture and long-range spatial structure, providing coherent segmentation of spills with diverse shapes and sizes. We employ a compound loss configuration (JointLoss = BCE + Dice + SSIM) designed to address class imbalance and boundary diffuseness common in SAR oil spill imagery. To rigorously validate the model, we evaluate it across four datasets spanning two frequency bands (C and L), two polarizations (VV, HH), and spatial resolutions from 5 m to 150 m: MKLab (Sentinel-1), CIMAT (ENVISAT-ASAR), and the SOS dataset with ALOS-PALSAR and Sentinel-1 subsets. Results show that JointLoss consistently improves segmentation quality over standard BCE, with average gains of +1.8% IoU, +1.2% F1, and +3.5% Recall. The model achieves competitive or superior performance compared to published baselines where available. Additional cross-domain experiments highlight likely contributors to the observed performance differences across sensors, providing insights for future domain-adaptation studies.

Keywords: SAR, oil spill detection, semantic segmentation, Swin Transformer, UPerNet, multi-sensor evaluation, compound loss.

I. INTRODUCTION

Oil spills and leaks in the ocean are among the main threats to marine and coastal ecosystems, causing ecological and socioeconomic damage that may take years to reverse [1]. Although the frequency of large-scale disasters has decreased since the 1970s—largely due to improved regulations and double-hull tanker designs—smaller incidents associated with exploration, transportation, and offshore operations continue to occur regularly. The International Tanker Owners Pollution Federation (ITOPF) estimates that routine operational discharges and minor spills account for a substantial portion of annual petroleum input into the oceans, maintaining the need for effective surveillance and early-response strategies. The detection and monitoring of oil spills at sea requires sensors capable of covering vast ocean areas with sufficient temporal and spatial resolution. Optical satellite sensors, while useful under clear-sky conditions, are severely limited by cloud cover, nighttime operation, and sun-glint effects. In contrast, Synthetic Aperture Radar (SAR) sensors have proven particularly suitable for maritime monitoring due to their ability to operate day and night, regardless of weather conditions [2], [3]. Their physical principle, based on detecting changes in sea surface roughness via microwave backscatter, enables identification of the damping effect that hydrocarbon films exert on short gravity-capillary waves at the water surface. The interaction between microwaves and the ocean surface is governed by Bragg scattering theory [2]: when an oil film is present, it suppresses short gravity-capillary waves by increasing surface viscosity, producing a localized reduction in radar backscatter visible as a dark region in the SAR image. However, SAR-based oil

spill detection faces several technical challenges: (i) multiplicative speckle noise inherent to coherent imaging systems, (ii) look-alike phenomena (biogenic films, low-wind zones, rain cells, internal waves) that produce similar dark features, and (iii) variability across sensors in terms of frequency band, polarization, incidence angle, and spatial resolution [2]. Classical detection methods based on thresholding and machine learning with handcrafted features achieved limited robustness, with false positive rates reaching 50–70% [2]. Deep learning approaches—particularly CNN-based architectures such as U-Net and DeepLabV3+—have substantially improved performance [4], and, more recently, Vision Transformers have demonstrated greater structural coherence through self-attention mechanisms [5]. Despite the progress in deep learning for SAR oil spill detection, most existing methods are evaluated on a single dataset from a specific sensor, limiting our understanding of their generalization capabilities. Real-world operational scenarios involve multiple satellite platforms with different frequency bands (C-band at 5.4 GHz, L-band at 1.3 GHz), polarization modes (VV, HH, VH, HV), and spatial resolutions ranging from 1 m to 150 m. A model that performs well on Sentinel-1 C-band imagery may fail when applied to ALOS-PALSAR L-band data, due to differences in the physical interaction with the oil film and the underlying sea state. In this work, we adapt and evaluate a Swin Transformer backbone coupled with UPerNet for oil spill segmentation in single-channel SAR imagery. The Swin Transformer's hierarchical shifted-window attention provides an effective mechanism for capturing both local speckle patterns and the large-scale spatial structure of oil slicks. At the same time, UPerNet enables multi-scale feature fusion for precise boundary delineation. Our contributions are:

Adaptation of Swin-B + UPerNet pretrained on ADE20K to single-channel SAR input via filter averaging, demonstrating that a general-purpose vision architecture achieves competitive or superior performance for oil spill segmentation with minimal domain-specific modifications.

A compound loss function (BCE + Dice + SSIM) that consistently improves mask completeness and boundary coherence across all evaluated datasets, with particular effectiveness for weathered or diffuse spills.

Rigorous validation across four datasets spanning C-band and L-band SAR, with resolutions from 5 m to 150 m, demonstrating stable performance across diverse sensor configurations. Complementary cross-domain experiments identify the main sources of the domain gap for future work.

II. RELATED WORK

1. Classical Methods

Traditional oil spill detection in SAR imagery follows a three-stage pipeline: (1) dark spot detection via adaptive thresholding,

(2) feature extraction using geometric, textural, and contextual attributes, and (3) classification using machine learning [2]. While effective under controlled conditions, these methods require manual tuning for each sensor configuration, with false alarm rates reaching 50–70% due to look-alike phenomena that can account for up to 90% of detected dark spots [2]. Recent reviews [3] confirm that the field is moving toward deep learning, while cross-sensor generalization remains an open question.

2. CNN-Based Methods

Krestenitis et al. [4] published the first large public dataset with over 2,000 annotated SAR images from Sentinel-1. They benchmarked several CNN architectures for multi-class segmentation (oil spill, look-alike, ship, land,

sea). U-Net achieved an mIoU of 64.97%, and DeepLabV3+ reached 65.06%, significantly outperforming classical methods. The dataset introduced a standardized evaluation protocol that enabled fair comparison across methods. Subsequent improvements include multi-channel inputs integrating wind speed maps and auxiliary layers [6] (IoU 97.17%). Although these multi-channel approaches report high performance, they rely on auxiliary information and are therefore not directly comparable to single-channel SAR-only models. Li et al. [7] proposed multi-scale feature fusion modules (mIoU 80%) that address the varying sizes of oil spills, from small localized leaks to massive slicks extending hundreds of kilometers. Zhu et al. [8] introduced boundary-supervised networks with attention blocks (mIoU 83%, F1 87%), addressing the challenge of delineating diffuse spill boundaries. Huang et al. [9] proposed a two-stage Faster R-CNN + U-Net++ framework for Sentinel-1, emphasizing detection followed by delineation. Despite these advances, CNN architectures exhibit inherent limitations in modeling long-range spatial dependencies due to their local receptive fields. Oil spills can extend over very large, irregularly shaped areas with variable internal intensities, requiring the model to integrate information across distant spatial locations. This limitation motivates the exploration of attention-based models capable of capturing global context.

3. Transformer-Based Approaches

The Vision Transformer (ViT) [5] introduced self-attention to image recognition, dividing images into fixed-size patches and processing them as sequences. While powerful, standard ViT has quadratic complexity in sequence length and lacks a built-in multi-scale hierarchy, making it computationally expensive for dense prediction tasks such as segmentation. Dehghani-Dehcheshmeh et al. [10] proposed hybrid CNN–Transformer architectures (DL-ViT variants) that combine convolutional feature extraction with Transformer-based global attention. Their best model achieved mIoU of 64.59% and F1 of 78.48% on the MKLab dataset, demonstrating that attention mechanisms improve spatial consistency in segmentation outputs. Basit et al. [11] explored CNN–ViT combinations with gradient profile loss, achieving an mIoU of 78.45%, where the gradient profile loss was specifically designed to improve edge delineation. Li et al. [7] compared their multi-scale CNN against TransUNet and Swin-UNet, reporting competitive results that highlight the need for hierarchical multi-scale attention in SAR applications. More recently, Sai et al. [12] introduced Ocean Shield, a Transformer-based approach for marine oil spill detection from SAR imagery, reflecting the growing interest in attention-based models for this application. Sun et al. [13] proposed MAF-UFormer, a U-shaped Transformer incorporating multi-scale alignment and fusion mechanisms for SAR oil spill segmentation, and evaluated it on both the Sentinel-1 and PALSAR subsets of the SOS dataset. These developments reinforce the relevance of hierarchical and multi-scale attention for representing oil spills with heterogeneous sizes, textures, and boundary characteristics. The Swin Transformer [14] addresses the computational limitations of standard ViT through shifted-window self-attention, achieving linear complexity with respect to image size while maintaining the ability to capture long-range dependencies through cross-window connections. Its hierarchical structure naturally produces multi-scale feature maps well-suited for dense prediction, making it an attractive backbone for segmentation tasks. Our work combines Swin-B with UPerNet and validates it across multiple SAR sensors, addressing the single-dataset limitation common in prior evaluations. More broadly, foundation models such as SAM [15] have attracted attention for zero-shot segmentation, though their application to single-channel SAR with severe speckle and look-alike phenomena remains unexplored. Domain adaptation techniques represent complementary strategies to improve cross-sensor transfer and constitute a relevant direction for future work.

III. METHODOLOGY

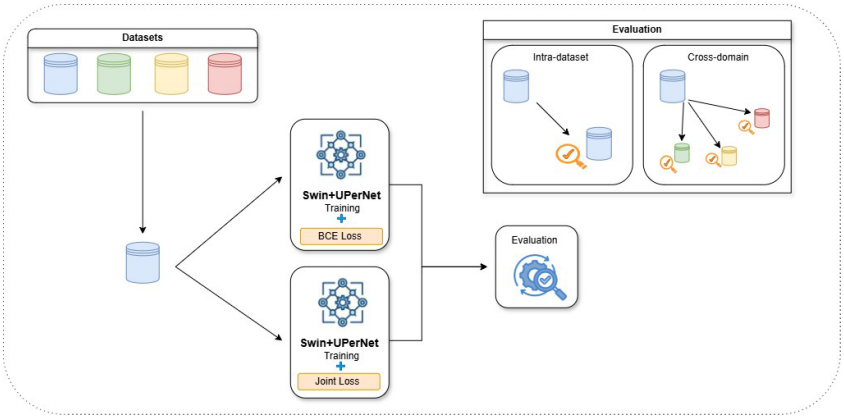


Figure 1 presents an overview of the proposed methodology, which consists of four main stages: data preparation and normalization, SAR-adapted model architecture, compound loss training, and multi-sensor evaluation.

Figure 1. Overview of the proposed multi-sensor oil spill segmentation methodology. The pipeline includes data preparation from multiple SAR sensors, model adaptation for single-channel input, training with compound loss, and cross-domain evaluation.

1. Model Architecture

We employ Swin-B [14] as the backbone encoder coupled with UPerNet [16] as the segmentation head. The Swin Transformer introduces two key innovations over the standard Vision Transformer: (1) hierarchical feature maps with progressive down-sampling, producing representations at multiple scales analogous to CNN feature pyramids, and (2) shifted-window self-attention that restricts computation to local windows while enabling cross-window interaction through alternating shift patterns.

a) Swin-B Encoder. Swin-B constructs hierarchical feature maps through four stages with channel dimensions {128, 256, 512, 1024} and block counts {2, 2, 18, 2}. Each stage consists of multiple Swin Transformer blocks, where each block applies Layer Normalization, Window-based Multi-Head Self-Attention (W-MSA or SW-MSA), another Layer Normalization, and a 2-layer MLP with GELU activation.

The window-based self-attention computes attention within non-overlapping local windows of size $M \times M$ (default $M = 7$):

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V,$$

M kare×dє $\mathbb{R}^{M^2 \times d}$ are the query, key, and value matrices, d is the head dimension, and B is a learnable relative position bias. The shifted-window mechanism in alternating layers displaces the window partition by $(\lfloor M/2 \rfloor, \lfloor M/2 \rfloor)$ pixels, enabling cross-window connections without increasing computational cost.

Between stages, patch-merging layers concatenate features from 2x2 neighboring patches and apply a linear projection to halve the spatial resolution while doubling the number of channels. These yield feature maps at

resolutions $\left\{\frac{H}{4}\times\frac{W}{4},\frac{H}{8}\times\frac{W}{8},\frac{H}{16}\times\frac{W}{16},\frac{H}{32}\times\frac{W}{32}\right\}$ with corresponding channel dimensions, providing the multi-scale representations required for dense prediction.

b) Adaptation to Single-Channel SAR. The original patch embedding layer is designed for 3-channel RGB input, using a convolutional layer with a 4×4 kernel and a stride of 4. For single-channel SAR images, we adapt the pretrained filters by averaging across the channel dimension:

$$W_{SAR} = \frac{1}{3}\sum_{c=1}^{3} W_{RGB}[:, c, :, :],$$

producing a single-channel input projection $W_{SAR} \in \mathbb{R}^{124\times1\times4\times4}$ while preserving the 128-dimensional output. This strategy avoids architectural modifications and enables effective transfer learning from ImageNet-22K and ADE20K weights [17]. An alternative approach is to duplicate the SAR channel three times to match the original RGB input; however, filter averaging preserves an explicitly single-channel projection and avoids unnecessary memory overhead.

c) UPerNet Decoder. UPerNet [16] fuses the multi-scale encoder outputs through a Pyramid Pooling Module (PPM) applied to the deepest feature map and a Feature Pyramid Network (FPN) that merges all four stages via top-down lateral connections, producing aligned feature maps at ¼ resolution. The final prediction head applies a 3x3 convolution and bilinear up-sampling to the original input at the original input resolution. Figure 2 illustrates the complete architecture.

2. Loss Functions

The choice of loss function significantly impacts segmentation quality in SAR imagery, where class imbalance (ocean pixels vastly outnumber spill pixels) and boundary ambiguity are prevalent. We compare two training configurations:

BCE (baseline): Standard binary cross-entropy providing pixel-wise supervision:

$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[\bar{y}_i log(y_i) + (1-\bar{y}_i)log(1-y_i)]$,Wh
$i. is\ the\ predicted\ probability\ for\ pixel\ i, y$ çubuk alt $i.$, $\bar{y}_i$ is the ground truth label, and N is the total number of pixels. BCE treats each pixel independently, which can lead to fragmented predictions when the network is uncertain about contiguous regions.

JointLoss (compound): A weighted combination addressing complementary aspects of SAR segmentation:

$\mathcal{L}$ = LBCE + β LSSIM + γ LDice,

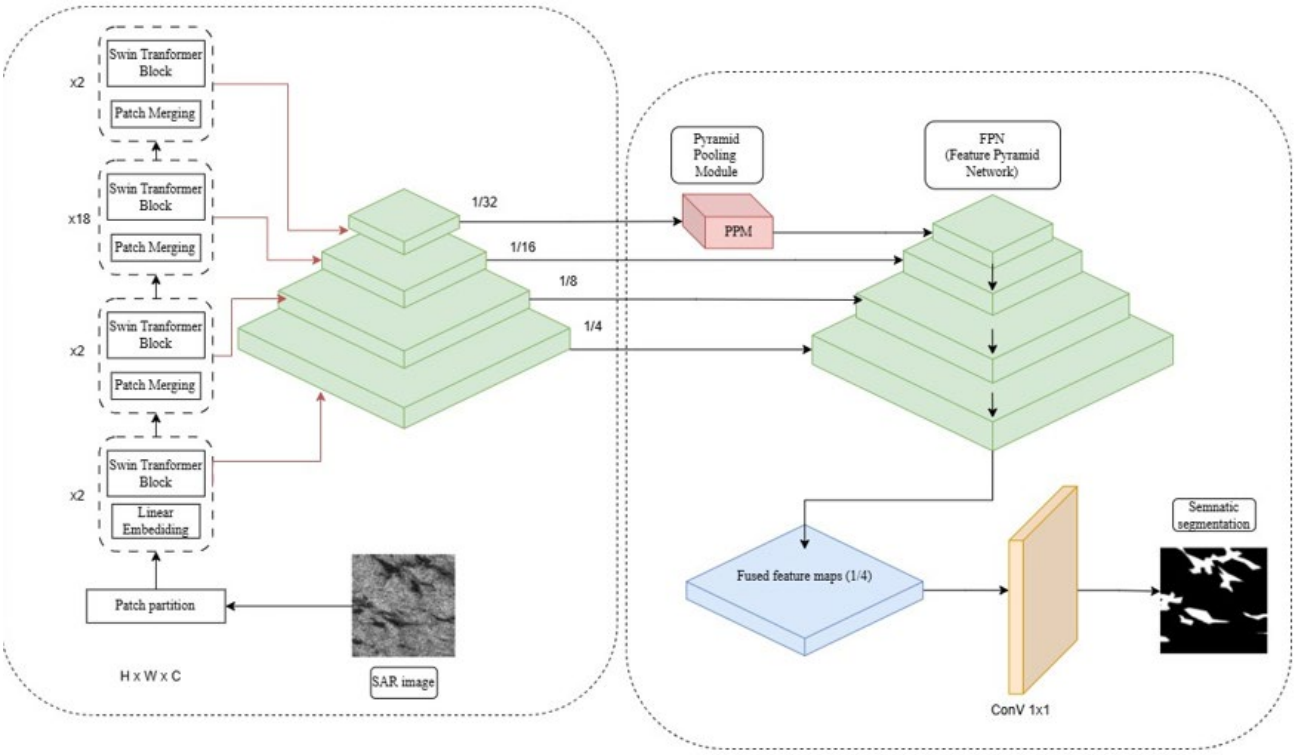


Figure 2. Detailed architecture of the Swin-B + UPerNet model for SAR oil spill segmentation, showing the hierarchical encoder stages, shifted-window attention mechanism, and the UPerNet decoder with PPM and FPN modules.

The Structural Similarity Loss is defined as:

$$L_{SSIM} = 1 - SSIM(y, \bar{y}) = 1 - \frac{(2\mu_y\mu_{\bar{y}} + c_1)(2\sigma_{y\bar{y}} + c_2)}{(\mu_y^2 + \mu_{\bar{y}}^2 + c_1)(\sigma_y^2 + \sigma_{\bar{y}}^2 + c_2)}$$

where μ and σ denote the local mean and variance computed over a sliding window, $\sigma_{y\bar{y}}$ is the local covariance, and $c1$, $c2$ are stability constants. SSIM enforces structural coherence between the prediction and the ground truth at the local patch level, which is particularly beneficial for oil spills with diffuse boundaries resulting from weathering and emulsification. Unlike pixel-wise losses, SSIM captures the structural pattern of the segmentation mask, penalizing fragmented predictions that are structurally inconsistent with the ground truth.

The Dice Loss directly optimizes the overlap metric:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \bar{y}_i}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \bar{y}_i^2}$$

which mitigates class imbalance by focusing on the foreground overlap rather than treating all pixels equally. This is critical in oil spill segmentation, where spill pixels often constitute less than 10% of the image area.

. The weights $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 0.6$ were determined via preliminary grid search on internal validation splits, explring configuration in the r ange $\alpha, \beta \in \{0.3, 0.5, 1.0\}$ and $\gamma \in \{0.3, 0.6, 0.8\}$. The selected combination provided the best trade-off between Recall and Precision on the validation set of the SOS-PALSAR dataset. These weights were kept fixed across all subsequent experiments to ensure fair comparison and avoid per-dataset overfitting. The rationale for these weights reflects the complementary roles: BCE provides a gradient signal for all pixels (including easy negatives), SSIM ensures structural consistency at the patch level, and Dice focuses the optimization on the minority class. A slightly higher weight for Dice ($\gamma = 0.6$) compensates for the severe class imbalance typical of SAR oil spill imagery, where spill pixels often account for less than 10% of the patch area.

3. Datasets

Four datasets spanning different sensors, bands, spatial resolutions, and geographic regions are used (Table 1). The SAR frequency bands employed in this study correspond to two distinct portions of the microwave pectru: C-band at a frequeny of approxmately 405 GHz (waelength $\lambda = 5.6\ cm), whiband\ operates\ at\ approximately\ 1.270\ Gwavelength\ wavelength\ Hz$ ($\lambda = 3.6$ cm). These bands interact differently with the ocean surface and oil films: C-band is generally more sensitive to short gravity-capillary waves, whereas L-band responds to longer surface structures and may be less sensitive to small-scale roughness variations. These differences can affect the apparent contrast, texture, and boundary definition of oil slicks. Regarding polarization, VV (Vertical-Vertical) denotes vertically polarized transmission and reception, the standard co-polarization mode for ocean monitoring due to its enhanced sensitivity to Bragg scattering from the sea surface. HH (Horizontal-Horizontal) denotes horizontally polarized transmission and reception, which provides complementary information about surface roughness and has shown effectiveness in detecting oil films at lower incidence angles. Spatial resolution varies substantially across datasets, ranging from 5 m (Sentinel-1A azimuth) to 150 m (ENVISAT-ASAR WSM mode). This diversity is central to our evaluation, as resolution determines the level of detail captured within oil slicks and directly affects the model's ability to delineate spill boundaries. The CIMAT dataset is particularly interesting in this regard, as it combines patches from two acquisition modes (IMP at 15 m and WSM at 150 m), introducing intra-dataset resolution variability.

Table 1. Summary of datasets used in this study. Band frequencies: C-band ≅ 5.405 GHz ($\lambda \cong 5.6$ cm), L-band ≅ 1.270 GHz ($\lambda \cong 23.6$ cm). Polarization: VV = vertical transmit/vertical receive; HH = horizontal transmit/horizontal receive.

Dataset	Sensor	Res. (m)	Band (freq.)	Pol.	Patch	Train	Test
MKLab [4]	Sentinel-1	10	C (5.405 GHz)	VV	160×160	3900	420
CIMAT [6]	ENVISAT-ASAR	15 / 150	C (5.331 GHz)	VV	224×224	2508	628
SOS-PALSAR [8]	ALOS-PALSAR	12.5	L (1.270 GHz)	HH	256×256	3101	776
SOS-S1 [8]	Sentinel-1A	5×20	C (5.405 GHz)	VV	256×256	3354	839

Each dataset preserves its original preprocessing pipeline and train/test split to ensure comparability with published baselines. All patches contain at least one positive pixel (oil spill). The datasets cover different geographic regions (Gulf of Mexico, Persian Gulf, Black Sea, global coastal areas), acquisition conditions, and spatial resolutions spanning over one order of magnitude (5–150 m), providing a comprehensive testbed for evaluating model robustness across the operational variability encountered in real SAR monitoring systems.

4. Training Protocol

All experiments use the AdamW optimizer with an initial learning rate of 1x10-4 and weight decay of 1x10-2, learning rate reduction on plateau (factor 0.5, patience 3), early stopping (patience 5), batch size 16, and a maximum of 30 epochs. Training was performed on an NVIDIA RTX 3090 (24 GB VRAM). Hyperparameters were held constant across all datasets and cross-domain experiments to ensure a fair comparison. Data augmentation: To improve generalization and reduce overfitting, we apply random geometric transformations during training: horizontal and vertical flips (probability 0.5 each), random rotations (±15◦), and random scaling (factor 0.8-1.2). No photometric augmentations (contrast, brightness) are applied because SAR intensity values carry physical meaning related to surface backscatter coefficients, and arbitrary photometric transformations could disrupt the radiometric relationships the model needs to learn. Preprocessing: Each dataset maintains its original preprocessing pipeline to ensure comparability with published results. MKLab and SOS datasets provide patches already extracted and normalized; CIMAT applies patch-level min-max normalization to [0, 1]. All images are resized to the network input size (224 × 224) using bilinear interpolation for images and nearest-neighbor for masks. Model initialization: The Swin-B backbone is initialized from weights pretrained on ImageNet-22K and fine-tuned on ADE20K for semantic segmentation. The UPerNet decoder is also initialized from ADE20K pretraining. Only the patch embedding layer is modified for single-channel input as described in Section III. All layers are fine-tuned end-to-end with the same learning rate. Model complexity and inference cost: The complete Swin-B + UPerNet architecture contains 123.5M parameters and requires 55.9 GFLOPs per forward pass for a 224 × 224 single-channel input. On an NVIDIA TITAN RTX GPU, the average inference time is 42.2ms per patch (batch size 1), corresponding to an approximate throughput of 24 patches/second, with a peak GPU memory footprint of 2.0 GB during inference.

IV. RESULTS AND DISCUSSION

1. Intra-Dataset Performance

Table 2 presents the test-set performance across all datasets, comparing our model against the best published baselines.

Table 2. Test results across datasets. Best results per dataset in bold. Baseline results are taken from their respective original publications and were not reproduced under identical splits, hardware, or evaluation protocol; comparisons should therefore be interpreted as contextual references rather than a strictly controlled benchmark. All our metrics correspond to the oil-spill foreground class computed pixel-wise on the test set.

Dataset	Method	Acc.	Prec.	Rec.	F1	IoU
SOS-PALSAR	Multi-Map [7]	0.930	0.729	–	0.784	–
	Ours (BCE)	0.931	**0.777**	0.824	0.800	0.666
	Ours (JointLoss)	**0.932**	0.765	**0.863**	**0.811**	**0.682**
SOS-S1	Multi-Map [7]	0.879	0.769	–	0.815	–
	Ours (BCE)	**0.898**	**0.865**	0.837	0.850	0.740
	Ours (JointLoss)	0.895	0.806	**0.919**	**0.859**	**0.753**
MKLab	DL-ViT-2 [10]	0.965	**0.861**	0.721	**0.785**	**0.646**
	Ours (BCE)	0.961	0.748	0.787	0.767	0.623
	Ours (JointLoss)	**0.967**	0.768	**0.795**	0.781	0.641

CIMAT	Ours (BCE)	0.875	0.832	0.897	0.863	0.759
	Ours (JointLoss)	**0.890**	**0.851**	**0.909**	**0.879**	**0.785**

On SOS-PALSAR, JointLoss improves Recall by +3.9 percentage points and IoU by +1.6 points over BCE, producing more complete masks for weathered spills. These differences may reflect the distinct imaging characteristics of L-band SAR, whose longer wavelength can produce smoother spatial responses and different sensitivity to surface roughness variations, favoring more continuous predictions in this dataset. On SOS-Sentinel-1, the model surpasses Multi-Map [7] with an F1 of 0.859 vs. 0.815. JointLoss achieves a substantial Recall improvement (+8.3 pp over BCE), indicating more complete spill coverage at the cost of slightly lower Precision, consistent with the loss design prioritizing structural completeness. C-band provides higher surface contrast but is more sensitive to local wind conditions and look-alike interference. On MKLab, performance is comparable to the best hybrid CNN-Transformer (DL-ViT-2: F1 0.785 vs. ours 0.781), with notably higher Recall (0.795 vs. 0.721), indicating better spill coverage. The slightly lower Precision is attributable to imprecise boundary annotations in the reference masks, where correctly detected spill regions beyond the annotated contour are penalized as false positives.

CIMAT shows the strongest absolute performance (F1 0.879, IoU 0.785), possibly due to differences in preprocessing, acquisition mode, annotation consistency, and coarser WSM resolution, which may reduce high-frequency speckle variability.

Across all four datasets, JointLoss improves Recall, F1, and IoU over BCE, with average gains of +1.8% IoU, +1.2% F1, and +3.5% Recall, although the increased mask completeness reduces Precision in some cases.

2. Cross-Domain Evaluation

Complementary cross-domain experiments were conducted to assess how the adapted Swin-B + UPerNet model behaves when transferred across datasets and sensors. These experiments are not intended to establish a fully optimized domain adaptation strategy, but rather to identify the main factors affecting transferability in SAR oil spill segmentation.

Table 3 summarizes the transfer results for both loss configurations, reporting F1 and IoU when training on one dataset and testing on others.

Table 3. Cross-domain evaluation: F1 / IoU. Rows = training set; columns = test set. (a) BCE loss; (b) JointLoss.

(a) BCE

Train \ Test	CIMAT	MKLab	PALSAR	S1
CIMAT	–	0.3040 / 0.1792	0.3262 / 0.1949	0.7267 / 0.5707
MKLab	0.1174 / 0.0624	–	0.6811 / 0.5164	0.3589 / 0.2187
PALSAR	0.3148 / 0.1868	0.5209 / 0.3521	–	0.8199 / 0.6947
S1	0.1888 / 0.1042	0.5142 / 0.3461	0.7699 / 0.6259	–

(b) JointLoss

Train \ Test	CIMAT	MKLab	PALSAR	S1
CIMAT	–	0.3139 / 0.1862	0.5260 / 0.3569	0.7066 / 0.5463
MKLab	0.0550 / 0.0283	–	0.6598 / 0.4923	0.3024 / 0.1781
PALSAR	0.5218 / 0.3530	0.5218 / 0.3530	–	0.8082 / 0.6781
S1	0.2243 / 0.1263	0.5331 / 0.3634	0.7777 / 0.6362	–

Key findings:

Transfers involving the two SOS subsets show the highest out-of-domain robustness. PALSAR→S1 achieves F1=0.808, while S1→PALSAR yields F1=0.778, suggesting that annotation consistency and related acquisition contexts facilitate cross-band transfer.

Band alignment does not guarantee optimal transfer: PALSAR→S1 (L→C, F1=0.808) outperforms CIMAT→S1 (C→C, F1=0.707), indicating that annotation consistency, preprocessing protocol, and spill morphology may contribute more than frequency match alone.

CIMAT, as a target, exhibits limited transfer. The best result comes from PALSAR→CIMAT (F1=0.522), while MKLab→CIMAT degenerates almost completely (F1=0.055). This failure likely stems from a combination of factors: (i) CIMAT applies patch-level min-max normalization that produces a distinct intensity distribution compared to the global normalization used in MKLab; (ii) CIMAT combines two acquisition modes (IMP at 15 m and WSM at 150 m), creating intra-dataset resolution variability absent from MKLab's uniform 10 m Sentinel-1 patches; (iii) annotation conventions differ—CIMAT includes diffuse spill boundaries typical of WSM's coarser pixel spacing, while MKLab annotations follow sharper contours. These compounding mismatches render the MKLab-trained features largely uninformative for CIMAT's target domain. Compared to BCE, JointLoss improves transfer in 8 of 12 out-of-domain pairs, with average gains of +2.6 pp in F1 and +12.9 pp in Recall. The improvement is most pronounced for transfers involving CIMAT and SOS-PALSAR as source or target. However, when MKLab is used as the source, JointLoss reduces F1 across all three target datasets, suggesting that the increased Recall-driven component of the compound loss amplifies prediction noise when the source-domain features are already poorly matched to the target distribution.

3. Qualitative Analysis

Figure 3 shows representative segmentation examples from SOS-PALSAR, comparing BCE and JointLoss predictions against ground truth. JointLoss produces more continuous masks, filling thin or interrupted regions without requiring overly sharp boundaries—a behavior consistent with the quantitative Recall improvement and physically plausible given the gradual transitions observed in weathered spills. Although visual differences between BCE and JointLoss predictions may appear subtle at display scale, they are operationally significant. At PALSAR's 12.5 m pixel spacing, even a few recovered pixels along spill boundaries correspond to tens of meters of additional detected coverage. The quantitative improvements reported in Table 2 (+3.9 pp Recall, +1.6 pp IoU) confirm that these subtle spatial differences translate into measurable segmentation gains. Similar qualitative patterns were observed on the CIMAT dataset, where JointLoss improved boundary delineation in coarser WSM imagery.

In both datasets, JointLoss is most beneficial for elongated, thin spills where BCE alone produces fragmented predictions with isolated positive pixels that are spatially inconsistent.

V. DISCUSSION

The results demonstrate the effectiveness of Swin-B + UPerNet for SAR oil spill segmentation across diverse sensor configurations:

Architecture suitability: The Swin Transformer's hierarchical shifted-window attention captures both local texture patterns and global spatial structure of oil slicks. Unlike standard CNNs, cross-window connections enable coherent segmentation of large, irregularly shaped spills without requiring explicit multi-scale input pipelines. Loss function design: JointLoss consistently improves Recall, F1, and IoU across all datasets, sometimes at the cost of lower Precision. The +3.5% average Recall gain is critical for operational monitoring, where underestimating the extent of a spill can have severe environmental consequences. Multi-sensor stability: The model achieves competitive results across all four datasets without any architecture modifications between sensors, confirming that large-scale pretraining with filter-averaging adaptation supports stable performance across C-band and L-band SAR with different resolutions and polarizations. Cross-domain insights: The transfer patterns suggest that preprocessing consistency, annotation quality, and spatial-resolution alignment are important contributors to the observed domain gap, whereas frequency-band matching alone is insufficient to explain cross-domain performance. Domain adaptation remains a promising direction for future work.

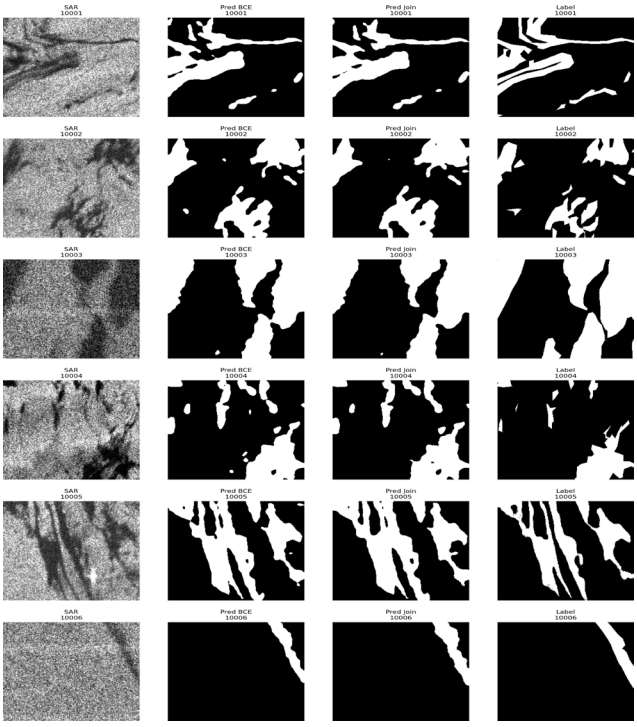


Figure 3. Segmentation examples on SOS-PALSAR (L-band, HH). Each row shows: SAR image, BCE prediction, JointLoss prediction, and ground truth (GT). JointLoss produces more complete masks with fewer fragmented regions.

VI. CONCLUSION

This work demonstrated that the Swin Transformer (Swin-B), coupled with UPerNet and adapted to single-channel SAR input, is an effective and versatile architecture for oil-spill semantic segmentation. The hierarchical shifted-window attention mechanism captures both local texture patterns and long-range spatial dependencies inherent to oil slick morphology, achieving competitive performance compared with published CNN and hybrid CNN–Transformer baselines where available.

The compound JointLoss (BCE + Dice + SSIM) consistently improves segmentation quality, with average gains of +1.8% IoU,

+1.2% F1, and +3.5% Recall over standard BCE training. The Recall improvement is particularly relevant for operational monitoring, where detecting the full extent of a spill takes priority over minimizing false positives. The combination of overlap-based and structural loss components is particularly beneficial for elongated, weathered, or fragmented spills, for which BCE alone tends to produce discontinuous masks. The model was rigorously validated across four datasets spanning C-band (Sentinel-1, ENVISAT-ASAR) and L-band (ALOS-PALSAR), with resolutions ranging from 5 m to 150 m and two polarization modes (VV, HH). This multi-sensor evaluation confirms the stability of the proposed approach under diverse acquisition conditions. Complementary cross-domain experiments suggest that differences in radiometric normalization, inconsistencies in annotation protocols, and resolution mismatches are important contributors to the observed domain gap. Several limitations should be acknowledged. First, because all evaluated patches contain at least one oil-spill pixel, the study does not quantify scene-level false alarm rates on entirely oil-free images or on scenes composed exclusively of look-alike phenomena (biogenic slicks, low-wind zones, rain cells). Operational deployment requires measuring specifically these conditions to ensure the model does not trigger excessive false positives in routine surveillance. Second, the reported baseline comparisons rely on results from the original publications rather than reproduction under identical experimental conditions, so performance differences should be interpreted with appropriate caution. Third, the loss weights were selected via a limited grid search without a full sensitivity analysis. Future work includes: (1) domain adaptation and fine-tuning strategies to improve cross-sensor transfer with minimal target labels, including recent foundation-model approaches such as the Segment Anything Model (SAM) adapted for remote sensing [15] and domain-adaptation frameworks for SAR imagery; (2) standardized radiometric normalization based on calibrated sigma-nought values; (3) evaluation of lightweight variants (Swin-T, SegFormer-B0) for resource-constrained deployment; (4) a comprehensive multi-model benchmark under the same multi-sensor protocol to establish standardized evaluation practices; and (5) systematic evaluation on oil-free and look-alike-only scenes to quantify operational false alarm rates.

REFERENCES

[1]. R. Al-Ruzouq, M. B. A. Gibril, A. Shanableh, A. Kais, O. Hamed, S. Al-Mansoori, and M. A. Khalil, “Sensors, features, and machine learning for oil spill detection and monitoring: A review,” Remote Sensing, vol. 12, no. 20, p. 3338, 2020.

[2]. W. Alpers, B. Holt, and K. Zeng, “Oil spill detection by imaging radars: Challenges and pitfalls,” Remote Sensing of Environment, vol. 201, pp. 133–147, 2017.

[3]. S. Dong, J. Feng, Z. Gu, K. Yin, and Y. Long, “A review of artificial intelligence and remote sensing for marine oil spill detection, classification, and thickness estimation,” Remote Sensing, vol. 17, no 22, p. 3681, 2025.

[4]. M. Krestenitis, G. Orfanidis, K. Ioannidis, K. Avgerinakis, S. Vrochidis, and I. Kompatsiaris, “Oil spill identification from satellite images using deep neural networks,” Remote Sensing, vol. 11, no. 15, p. 1762, 2019.

[5]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and

[6]. N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” International Conference on Learning Representations (ICLR), 2021.

[7]. R. Hasimoto-Beltrán, M. Canul-Ku, G. Díaz-Méndez, and F. Ocampo-Torres, “Ocean oil spill detection from SAR images based on multi-channel deep learning semantic segmentation,” Marine Pollution Bulletin, vol. 188, p. 114651, 2023.

[8]. C. Li, Y. Yang, X. Yang, D. Chu, and W. Cao, "A novel multi-scale feature map fusion for oil spill detection of SAR remote sensing," Remote Sensing, vol. 16, no 10, p. 1684, 2024.
[9]. Q. Zhu, Y. Zhang, Z. Li, X. Yan, Q. Guan, Y. Zhong, L. Zhang, and D. Li, "Oil spill contextual and boundary-supervised detection network based on marine SAR images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–10, 2022.
[10]. X. Huang, B. Zhang, and W. Perrie, "A two-stage deep learning method for marine oil spill localization and segmentation from synthetic aperture radar images," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 12 315–12 327, 2025.
[11]. S. Dehghani-Dehcheshmeh, M. Akhoondzadeh, and S. Homayouni, "Oil spills detection from SAR Earth observations based on a hybrid CNN transformer networks," Marine Pollution Bulletin, vol. 190, p. 114834, 2023.
[12]. A. Basit, M. A. Siddique, M. K. Bhatti, and M. S. Sarfraz, "Comparison of CNNs and vision transformers-based hybrid models using gradient profile loss for classification of oil spills in SAR images," Remote Sensing, vol. 14, no. 9, p. 2085, 2022.
[13]. M. J. Sai, L. S. Reddy, and K. Abinaya, "Ocean shield: Transformer-based detection of marine oil spills from SAR images," in 2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2026, pp. 1045–1052.
[14]. C. Sun, D. Wang, M. Xu, S. Wei, S. Liu, and Z. Li, "MAF-UFormer: Oil spill detection in SAR images using multi-scale alignment and fusion u-shaped transformer network," Regional Studies in Marine Science, vol. 94, p. 104773, 2026.
[15]. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in
[16]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
[17]. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4015–4026.
[18]. T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," 2018.
[19]. I. Ahmad and S. Shin, "An approach to run pre-trained deep learning models on grayscale images," in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021, pp. 177–180.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

FedSLM: A Federated Learning Approach Based on Fine-Tuning a Small Language Model for Text Classification

Nhat-Thien Nguyen
Department of Information Technology
Quy Nhon University
Gia Lai, Vietnam
nguyennhatthien119@st.qnu.edu.vn

Quang-Hung Le
Department of Information Technology
Quy Nhon University
Gia Lai, Vietnam
lequanghung@qnu.edu.vn
ORCID: 0000-0002-4796-1664

Abstract
Text classification is a fundamental task in natural language processing, with applications including sentiment analysis, spam detection, and information retrieval. Deploying these systems in real-world application scenarios requires not only accuracy but also efficiency and the preservation of privacy. Centralized training of large language model-based approaches poses significant challenges related to data privacy, communication overhead, and inference latency. This paper investigates the use of fine-tuned small language models for text classification in federated learning environments. We propose FedSLM, a novel federated framework that combines small language models with parameter-efficient fine-tuning via LoRA (Low-Rank Adaptation). Our experiments on three benchmark datasets show that the proposed framework achieves an average accuracy of 91.85% while reducing the amount of data transmitted per communication round by more than 75 times. These results demonstrate that the proposed framework is an effective solution that balances accuracy, privacy protection, and resource efficiency, making it well-suited for federated learning systems in resource-constrained environments. The source code is publicly available at https://github.com/NguyenNhatThienQNU/FedSLM.

Keywords: Federated Learning, Small Language Models, LoRA, Parameter-Efficient Fine-Tuning, Text Classification.

I. INTRODUCTION

Text classification is a fundamental task in natural language processing (NLP), powering applications from sentiment analysis to news categorization and content moderation [1]. In these applications, the text is generated on users' own devices and carries sensitive information, so it cannot be freely collected on a central server for training. This has motivated the development of federated learning (FL) [2], a distributed paradigm that enables model training across clients without transferring raw data to a central server, thereby preserving privacy while supporting large-scale learning. According to the modern approach, FL has achieved impressive performance through large-scale language models [3]. However, the deployment process introduces challenges in computational cost and inference latency. Consequently, recent research has turned to small language models (SLMs) [4], [5], which retain much of the accuracy of larger models while being far lighter and deployable on constrained hardware. In this paper, we investigate whether such SLMs can support accurate text classification within FL while keeping communication and computation low. To adapt an SLM without the cost of full fine-tuning, we employ parameter-efficient fine-tuning (PEFT). In particular, LoRA [6] freezes the backbone and trains only a small set of low-rank matrices, so that each client exchanges just these matrices and the classification head rather than the entire model. To the best of our knowledge, this is among the first works to jointly exploit SLMs and LoRA for text classification in a federated, resource-constrained setting. The main contributions of this work are summarized as follows:

1. We propose FedSLM, a novel federated framework that combines SLMs with LoRA for text classification, where each client fine-tunes and transmits only the LoRA parameters and the classification head instead of the full backbone.
2. We investigate the effectiveness of the PEFT methods to reduce the computational cost while maintaining a competitive model performance.
3. We evaluate the proposed method on three benchmarks under a non-IID federated setting, comparing it with classical machine learning, centralized training, and a larger language model fully fine-tuned under federated learning, and analyze the trade-offs among classification accuracy, communication cost, and resource efficiency.
4. We demonstrate the potential of SLMs for federated text classification in resource-constrained environments.

The rest of the paper is organized as follows. Section II reviews some related work. Section III formalizes the problem and details the proposed method. Section IV presents the experimental setup and results. Finally, Section V concludes and outlines future work.

II. RELATED WORK

Text classification has evolved from early machine learning methods to deep neural architectures. Convolutional models for sentence-level and character-level classification [1] provided strong early baselines, and the Transformer architecture [7] later reshaped the field. Pre-trained models such as BERT [8] and RoBERTa [9] became the de facto backbone for text classification by learning contextualized representations. These models, however, are typically trained in a centralized manner, storing all data on a single server, which raises data privacy concerns and hinders deployment across distributed devices. To address the aforementioned concerns, FL [2] shifted the paradigm from centralized training toward collaborative on-device learning, in which raw data never leaves the client. Building on FedAvg, optimization methods such as FedProx [10] improved convergence under non-IID data, while general frameworks such as Flower [11] and benchmarks such as FedNLP [12] extended FL to a range of NLP tasks [13]. Applying FL to large language models (LLMs), however, reintroduces heavy computation and per-round communication [3], thereby limiting their use in resource-constrained environments. In recent years, SLMs have emerged as a potential alternative to large models due to their low inference latency and cost-effectiveness [14], [15]. Distilled models such as DistilBERT [5] and TinyBERT [16] retain competitive accuracy within a compact parameter space [4], and remain effective even in industrial text classification settings [17]. Nevertheless, fully fine-tuning an SLM on every client still incurs substantial computation and communication in each federated round, and aggressively compressed models may sacrifice representational capacity. At the same time, PEFT offers a complementary approach to reducing this cost by updating only a small subset of parameters. Adapter-based tuning [18], notably LoRA [6], has become a standard PEFT technique [19]. Several studies have introduced LoRA into the federated setting, such as FedEx-LoRA [20] and FLoRA [21], to reduce communication overhead. A trend documented in recent surveys [22] and extended to multilingual modeling by FedLFC [23]. These efforts, however, largely target LLMs or text-generation tasks, leaving the joint use of SLMs and LoRA for federated text classification underexplored. This paper builds upon these advancements by integrating SLMs with LoRA to bridge the gap between resource efficiency and classification accuracy in resource-constrained federated environments.

III. PROPOSED METHOD

This section presents the proposed FedSLM framework. We first formalize the problem and describe the overall architecture. Then we introduce the framework's main components: the SLM backbone, the resource-efficient LoRA-based fine-tuning strategy, and the FedAvg-based federated optimization mechanism.

1. Problem Definition

Federated text classification is defined as the process of training a global classification model from data distributed across multiple clients without sharing their local data. Assume a system of K clients, where the k-th client holds a llocal dataset D_k = { (x_i, y_i) }, in which x_i is the input text and y_i e corresponding label. The objective is to find the model parameters θ that are optimal over the entire distributed dataset:

$$\min_{\theta} F(\theta) = \sum_{k=1}^{K} \frac{n_k}{N} F_k(\theta), \tag{1}$$

Here, $F_k(\theta)$I have the local loss function of the $k$the client, n_k is the number of samples on that client, and $N = \sum_{k=1}^{K} n_k$ is the total number of samples across the whole system [2].

The problem addressed here is to design an FL method that maintains text classification effectiveness while reducing computational and communication costs under heterogeneous data.

2. System Architecture

Figure 1 illustrates the overall architecture of FedSLM for text classification in the FL setting. The system comprises three main components. At the top is the central server, which hosts the global model comprising the SLM backbone, the LoRA parameters, and the classification head. In the middle are K clients, each holding its own text dataset and performing local LoRA fine-tuning. At the bottom is the FedAvg module, which receives parameters from clients to update the global model.

The dataset of the client k is denoted by the uppercase D subscript k, where n_k n_k is the number of samples on the client. The data D_k is stored entirely on the local device and is never transmitted to the server. The workflow consists of five steps:

1. The server initializes the global model, comprising the SLM backbone, LoRA, and the classification head.
2. The server sends the global LoRA parameters and classification head to the K clients.
3. Each client keeps the backbone frozen and fine-tunes LoRA and the classification head on its own data.
4. Each client sends the updated LoRA parameters, the updated classification head, and its sample count n_k to the server.
5. The server aggregates the received parameters with FedAvg to form a new global model.

Step 1 is performed only once before training begins. Thereafter, steps 2–5 are repeated over the communication rounds until the model converges or the maximum number of rounds is reached.

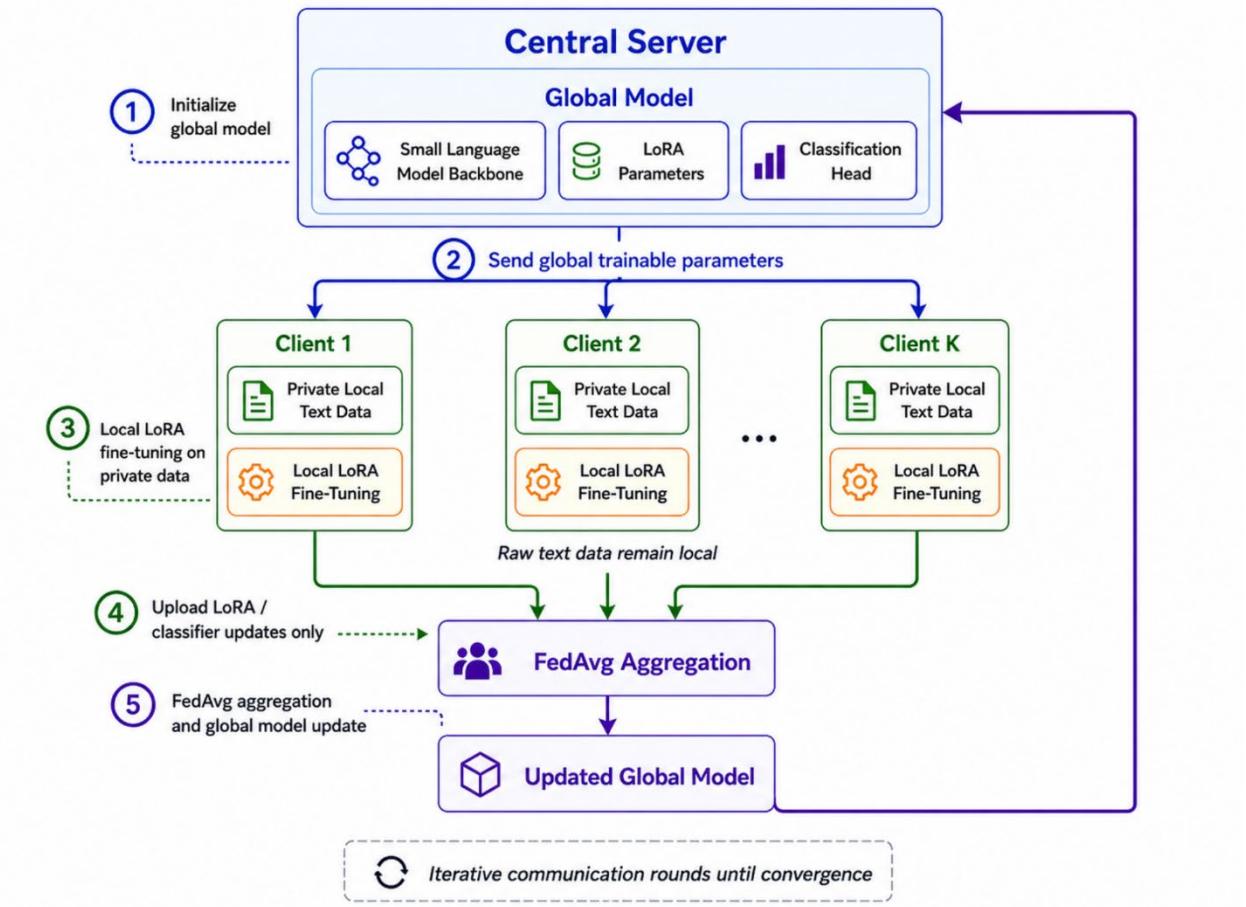


Figure 1. Overall workflow of the FedSLM framework.

3. Small Language Model Backbone

The SLM backbone is the central component of FedSLM, responsible for encoding the input text and producing the semantic representation used for classification. LoRA does not replace the backbone; instead, it adds low-rank updates to the targeted attention projections. The classification head then uses the representation of the [CLS] token to predict the label. In this study, DistilBERT [5] is chosen as the backbone. This model is smaller than BERT-base and is well-suited to clients in resource-constrained environments.

At step 1, the serout ×s the pre-trained weight set W_alt.W_b of the backbone. These weights are frozen throughout training. The server simultaneously attaches LoRA to the targeted attention projections and initializes a linear classification head. For the $j$the attention projection, with weights $W_{b,j} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA introduces two low-rank matrices:

$$A_j^{(0)} \in \mathbb{R}^{r \times d_{\text{in}}},\ B_j^{(0)} \in \mathbb{R}^{d_{\text{out}} \times r},\ r \ll \min(d_{\text{in}}, d_{\text{out}}). \tag{2}$$

classificaon head is denoted by (W_c, b_c), where $W_c \in \mathbb{R}^{C \times d}$ and $b_c \in \mathbb{R}^C$. Here, d is the representation size of the backbone and $C = |Y|$ is the number of classes. The global model at round t is represented as $M^{(t)} = \{W_b, \theta_L^{(t)}, \theta_c^{(t)}\}$, where $\theta_L^{(t)} = \{A_j^{(t)}, B_j^{(t)}\}_{j \in J}$ is the full set of LoRA parameters at the targeted attention

modules, and $\theta_c^{(t)} = \{W_c^{(t)}, b_c^{(t)}\}$ is the classification-head parameter set. The trainable global parameter set is denoted by $\theta^{(t)} = \{\theta_L^{(t)}, \theta_c^{(t)}\}$.

The backbone $is\ installed\ on\ or\ distributed\ to$the clients once before training begins. At step 2 of round t, the server sends only $\theta^{(t)} = \{A^{(t)}, B^{(t)}, W_c^{(t)}, b_c^{(t)}\}$ to each client $k = 1, \dots, K$. Since all clients use the same frozen backbone, W_b need not be retransmitted in each communication round.

4. Resource-Efficient Fine-Tuning

After receiving $\theta^{(t)}$, client k initializes its local parameter set $\theta_k^{(t,0)} = \theta^{(t)}$. The client combines these parameters with its local copy of the backbone to carry out step 3. LoRA is inserted directly into the backbone's attention projections and participates in producing the semantic representation. For the j The projection at client k, LoRA produces the update

$$\Delta W_{j,k} = \frac{\alpha_{\text{LoRA}}}{r} B_{j,k} A_{j,k}, \qquad W_{j,k} = W_{b,j} + \Delta W_{j,k}, \tag{3}$$

where r is the simge are $W_{b,j}$ remain unchanged; only the two matrices $A_{j,k}$ and $B_{j,k}$ are updated.

For each sample $x_i^{(k)}$, the client produces the semantic representation

$$h_i^{(k)} = \text{Encoder}_{\text{SLM}}\left(x_i^{(k)}; W_b, \theta_{L,k}\right).$$

The representation of the [CLS] token is fed into the local classification head. The logit vector and the predicted probability distribution are computed as

$$z_i^{(k)} = W_{c,k}\, h_{i,[\text{CLS}]}^{(k)} + b_{c,k}, \qquad \hat{p}_i^{(k)} = \text{Softmax}\left(z_i^{(k)}\right). \tag{4}$$

Client k minimizes the cross-entropy loss

$$F_k(\theta_k) = -\frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{c=1}^{C} \mathbf{1}\left[y_i^{(k)} = c\right] \log \hat{p}_{i,c}^{(k)}, \tag{5}$$

where $\hat{p}_{i,c}^{(k)}$ is the probability that sample i belongs to class c, and $\mathbf{1}[\cdot]$ is the indicator function.

AdamW is used to update the LoRA matrices $A_{j,k}, B_{j,k}$ and the classification-head parameters $W_{c,k}, b_{c,k}$. Gradients are not propagated to W_b, so the backbone remains unchanged throughout local training. After E local epochs, client k obtains

$$\theta_k^{(t,E)} = \{A_k^{(t,E)}, B_k^{(t,E)}, W_{c,k}^{(t,E)}, b_{c,k}^{(t,E)}\}.$$

At step 4, each client sends $\theta_k^{(t,E)}$ together with its sample count n_k to the central server. In Figure 1, the term “updates” refers to the LoRA and classification-head parameters that have been updated after local training. The text data D_k and the backbone W_b are not transmitted in this step.

5. Federated Optimization Strategy

Following step 4, the central server receives, from the K clients, the LoRA matrices, the classification-head parameters, and the sample counts $\{\left(\theta_1^{(t,E)}, n_1\right), \ldots, \left(\theta_K^{(t,E)}, n_K\right)\}$. After collecting all parameters, the server performs step 5 by applying FedAvg [2]:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} \theta_k^{(t,E)}, \qquad N = \sum_{k=1}^{K} n_k. \tag{6}$$

FedAvg is applied element-wise to the tensors $A_k, B_k, W_{c,k}$, and $b_{c,k}$. Clients with more data samples receive a larger contribution weight. The backbone W_b is neither uploaded nor aggregated, since it is frozen and identical across all clients.

After aggregation, the server obtains $\theta^{(t+1)} = \{A^{(t+1)}, B^{(t+1)}, W_c^{(t+1)}, b_c^{(t+1)}\}$. This parameter set is combined with the frozen backbone to form the updated global model $M^{(t+1)} = \{W_b, \theta^{(t+1)}\}$. In the next round, the server distributes $\theta^{(t+1)}$ to the clients at step 2. The distribution, local training, parameter upload, and aggregation are repeated until the validation loss no longer improves; communication rounds are reached.

IV. EXPERIMENTS

1. Experimental Setup

Datasets. To evaluate the generalization ability of the proposed method, we use three benchmark datasets: AG News [1], IMDB [24], and SST-2 [25]. These datasets cover both binary and multi-class classification, and include both sentence-level and document-level text, enabling us to assess the effectiveness of FedSLM across a range of text classification scenarios. Table 1 summarizes their statistics.

Table 1. Statistics of the benchmark datasets.

Dataset	Task	Classes	Train	Test
AG News	News topic classification	4	120,000	7,600
IMDB	Movie review sentiment classification	2	25,000	25,000
SST-2	Sentence sentiment classification	2	67,349	1,821

To simulate a realistic FL environment, the training set of each dataset is partitioned across 5 clients using a Dirichlet distribution with concentration parameter alpha = 0.5.. This setting produces a moderately non-IID scenario that reflects a common characteristic of FL systems, in which the data distribution across clients is heterogeneous but not overly imbalanced. In each communication round, the server distributes the global model to all clients. Each client performs one local training epoch on its own data following the procedure described in Section III. After training, only the LoRA parameters and the classification layers are sent back to the server for aggregation via FedAvg. At the same time, all text data is always stored locally on each client.

Baselines. To evaluate the effectiveness of FedSLM, we compare the proposed method against baselines that represent common approaches to text classification and FL. The baselines are divided into three groups:

- *Centralized SLM training.* DistilBERT [5] trained on centralized data with LoRA. This group is used to assess the impact of FL on classification effectiveness.
- *A larger language model under federated learning.* BERT-base [8] with full fine-tuning, an FL baseline with a larger backbone, used to compare classification effectiveness and communication cost against FedSLM.
- *Classical machine learning.* SVM with TF-IDF, which serves as a traditional machine learning baseline for text classification.

Evaluation Metrics. To assess both the prediction quality and the operating cost of the FL system, the metrics are divided into two groups: classification performance and resource efficiency. On all three datasets, the model is evaluated using accuracy (Acc) and macro-F1, which are regular evaluation metrics for text classification and information retrieval problems. The communication cost is determined by the size of the updated parameter set $|\Delta\theta|$, the number of clients K, and the number of communication rounds T:

$$C_{\text{total}} = 2 \times T \times K \times |\Delta\theta|. \quad (7)$$

In addition, we report the number of trainable parameters per client and the total training time, measured as the wall-clock time of a communication round.

Implementation Details. Each dataset is partitioned across 5 clients using a Dirichlet distribution (alpha =. 5). The system uses FedAvg for up to 20 communication rounds; each client performs one local epoch per round. DistilBERT is used as the SLM bbackbone LoRA is applied to the attention projections with $r = 8$, $\alpha_{\text{LoRA}} = 16$, and a dropout of 0.1. The model uses AdamW with early stopping at a patience of 3. The framework is implemented in Python with PyTorch, Hugging Face Transformers, and PEFT on Google Colab with a single NVIDIA T4 16 GB GPU. The same training program is used for both the centralized and the federated models to ensure a fair comparison.

V. RESULTS

Overall Performance. Table 2 compares our FedSLM against three groups of baselines:

- *Comparison with classical machine learning methods.* FedSLM outperforms the best classical machine learning method on all three datasets. The accuracy improvements are 8.83 points on SST-2, 0.85 points on IMDB, and 1.90 points on AG News. On average, the proposed method achieves 3.86 points higher accuracy and 3.89 points higher macro-F1. This result indicates that the Transformer encoder exploits contextual information more effectively than TF-IDF lexical features, especially for the short sentences in SST-2.
- *Comparison with centrally trained SLM.* FedSLM achieves 1.49 points higher on SST-2 and 0.26 points higher on IMDB, while being only 0.39 points lower on AG News. On average, FedSLM achieves 0.45 points higher accuracy and 0.47 points higher macro-F1 than the centralized model. This shows that moving to FL does not reduce the average performance, while the raw text data remains stored on the clients.
- *Comparison with a larger language model under federated learning.* The federated BERT-base with full fine-tuning achieves the highest average results, with 92.95% accuracy and 92.96% macro-F1. FedSLM achieves 91.85% and 91.86%, respectively, which is 1.10 points lower on both metrics. The gap on AG News is only 0.06 points on both metrics. Although BERT-base retains a performance advantage, the proposed method comes fairly close to the larger model at a substantially lower resource cost.

Table 2. Classification performance across three datasets.

Models	Setting	Tuning	Dataset	Acc (%)	Macro-F1 (%)
SVM	Classical	–	SST-2	81.65	81.60
			IMDB	90.09	90.09
			AG News	92.24	92.23
DistilBERT	Centralized	LoRA	SST-2	88.99	88.97
			IMDB	90.68	90.68
			AG News	94.53	94.50
DistilBERT	Federated	Full	SST-2	90.37	90.37
			IMDB	90.55	90.55
			AG News	94.09	94.10
BERT-base	Federated	Full	SST-2	93.00	93.00
			IMDB	91.66	91.66
			AG News	94.20	94.21
Ours (FedSLM)	Federated	LoRA	SST-2	90.48	90.48
			IMDB	90.94	90.94
			AG News	94.14	94.15

Resource Efficiency. Table 3 compares the number of trainable parameters and the update size of the three main FL configurations.

FedSLM updates only about 0.89 million parameters, corresponding to about 1.32% of the model's total parameters. Compared with the federated DistilBERT with full fine-tuning, the proposed method reduces the number of trainable parameters by about 98.67% and the update size by about 75×. Compared with the federated BERT-base with full fine-tuning, the number of updated parameters is reduced by about 99.19%, while the update size is reduced by about 123×. Although the average accuracy and macro-F1 are only 1.10 points lower than those of BERT-base, the proposed method requires exchanging only about 3.38 MB of parameters per round.

Training time requires more careful reading, as early stopping allows each configuration to run for a different number of rounds. Measured *per round*, FedSLM is the cheapest of the three federated configurations on every dataset: 82.4 s against 125.9 s for full fine-tuning of the same DistilBERT and 374.2 s for BERT-base on SST-2, 205.1 s against 232.0 s and 480.9 s on IMDB, and 294.7 s against 344.9 s and 898.5 s on AG News. Its higher *total* on IMDb (3691.9 s against 1856.2 s) therefore reflects convergence speed rather than per-round cost. By optimizing far fewer parameters, LoRA needed 18 rounds to reach its best checkpoint, compared to 8 for full fine-tuning. Compared with the federated BERT-base, FedSLM is faster overall on all three datasets. The advantage of LoRA thus lies in the number of updated parameters, the communication cost, and the per-

round compute, rather than in the total time to convergence. Overall, FedSLM achieves a suitable balance between classification performance and resource efficiency, indicating its potential for deployment in environments with limited memory and bandwidth.

Table 3. Comparison of resource efficiency.

Models	Setting	Tuning	Trainable params	Trained ratio	Update size	Training time (s)		
						SST-2	IMDB	AG News
DistilBERT	Federated	Full	66.96 M	100%	≈255.41 MB	503.4	1,856.2	3,449.5
BERT-base	Federated	Full	109.48 M	100%	≈417.65 MB	2,245.0	4,809.2	8,086.9
Ours (FedSLM)	Federated	LoRA	≈0.89 M	≈1.32%	≈3.38 MB	576.7	3,691.9	3,536.4

2. Discussion

The results show that FedSLM does not merely aim for the highest accuracy but also balances performance, privacy, and resource cost.

- *Trade-off among performance, privacy, and efficiency.* The proposed method achieves accuracy of 90.48%, 90.94%, and 94.14% on SST-2, IMDB, and AG News, respectively. Compared with the federated BERT-base with full fine-tuning, the largest drop is only 2.52 percentage points, while the number of parameters to be trained is reduced from 109.48 million to 0.89 million. The communication size per client in one round is also reduced from 417.65 MB to 3.38 MB. Notably, LoRA achieves results equal to or slightly better than those of the federated DistilBERT with full fine-tuning across all three datasets, showing that updating the entire model is unnecessary in the FL context. At the same time, the raw text data always remains stored locally, and only the LoRA parameters and the classifier are exchanged. However, this mechanism only provides privacy protection at the structural level; techniques such as differential privacy and secure aggregation still need to be added to defend against inference attacks based on the model updates.
- *Implications for real-world deployment.* With its small update size and low number of trainable parameters, FedSLM is well-suited to resource-constrained environments such as edge devices, mobile phones, and internal systems that handle sensitive data. The method can be applied to customer-feedback analysis, news classification, and the education, healthcare, and finance domains, where data cannot be centralized on a single server. Nevertheless, the current experiments are simulated using only five clients on an NVIDIA T4 GPU. Therefore, further evaluation on real devices, with a larger number of clients, and under unstable network conditions is needed before large-scale deployment.

VI. CONCLUSION

In this study, we proposed FedSLM, a novel framework that combines SLMs, LoRA, and FedAvg for federated text classification in resource-constrained environments. The framework freezes the backbone, updates only the LoRA parameters and the classification head, and then aggregates the updates on the server. The method was evaluated on three benchmark datasets with non-IID data distributed across five clients. It was compared against traditional machine learning methods, centralized training, and full fine-tuning in the FL setting. The experimental results confirm that combining SLMs with LoRA can maintain competitive classification performance without updating the entire model. This indicates that choosing the right parameters to fine-tune matters more than increasing model size, and shows that FedSLM achieves an effective balance among classification quality, data privacy, and resource cost. In doing so, the method opens up the possibility of deploying federated NLP systems on edge devices and in domains involving sensitive data.

In the future, we will extend the experiments to a larger number of clients, higher degrees of data heterogeneity, and realistic conditions such as heterogeneous devices, disconnections, or partial participation. Secure aggregation and differential privacy mechanisms will be integrated to strengthen security and robustness. In addition, we plan to multitask and use multilingual backbones, specialized adapters, and a suitable adapter-selection mechanism in multilingual settings.

ACKNOWLEDGEMENT

This research is conducted within the framework of science and technology projects at the institutional level of Quy Nhon University, Vietnam, under the project code T2026.923.07.

REFERENCES

[1] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 649–657.
[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th AISTATS, 2017, pp. 1273–1282.
[3] Y. Wu, C. Tian, J. Li, H. Sun, K. Tam, L. Li, and C. Xu, "A survey on federated fine-tuning of large language models," Transactions on Machine Learning Research (TMLR), 2026.
[4] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang, N. D. Lane, and M. Xu, "Small language models: Survey, measurements, and insights," arXiv preprint arXiv:2409.15790, 2024.
[5] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in Proc. ICLR, 2022.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. 3rd MLSys, 2020.

[11] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernández-Marqués, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," arXiv preprint arXiv:2007.14390, 2022.

[12] B. Y. Lin, C. He, Z. Zeng, H. Wang, Y. Huang, M. Soltanolkotabi, X. Ren, and S. Avestimehr, "FedNLP: Benchmarking federated learning methods for natural language processing tasks," in Findings of NAACL, 2022, pp. 157–175.

[13] M. Liu, S. Ho, M. Wang, L. Gao, Y. Jin, and H. Zhang, "Federated learning meets natural language processing: A survey," arXiv preprint arXiv:2107.12603, 2021.

[14] K.-T. Nguyen and Q.-H. Le, "Light-SQL: A Small Language Model Based Approach Using In-Context Learning for Natural Language to SQL," in Proc. International Conference on Computational Intelligence in Engineering Science, Springer, 2026, pp. 216–231.

[15] B. Nguyen and Q. Le, "A Novel Approach Based on Integrating Small Language Models and Retrieval-Augmented Generation for Medical Question Answering," Journal of Communications Software and Systems, vol. 22, no. 3, pp. 356–366, Jul. 2026, doi: 10.24138/jcomss-2026-0036.

[16] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," in Findings of EMNLP, 2020, pp. 4163–4174.

[17] L. Li, L. Sleem, N. Gentile, G. Nichil, and R. State, "Small language models in the real world: Insights from industrial text classification," in Proc. 63rd ACL (Industry Track), 2025, pp. 971–982.

[18] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in Proc. 36th ICML, 2019, pp. 2790–2799.

[19] Y. Mao, Y. Wang, Y. Dong, Y. Guo, W. Hu, Y. Shen, Y. Zhang, and J. Wang, "A survey on LoRA of large language models," Frontiers of Computer Science, vol. 19, art. 194606, 2025.

[20] R. Singhal, K. Ponkshe, and P. Vepakomma, "FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models," in Proc. 63rd ACL (Long Papers), 2025, pp. 1316–1336.

[21] Z. Wang, Z. Shen, Y. He, G. Sun, H. Wang, L. Lyu, and A. Li, "FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations," in Advances in Neural Information Processing Systems (NeurIPS), 2024.

[22] J. Bian, Y. Peng, L. Wang, Y. Huang, and J. Xu, "A survey on parameter-efficient fine-tuning for foundation models in federated learning," arXiv preprint arXiv:2504.21099, 2025.

[23] Z. Guo, Y. Zhang, Z. Zhang, Z. Xu, and I. King, "FedLFC: Towards efficient federated multilingual modeling with LoRA-based language family clustering," in Findings of NAACL, 2024, pp. 1519–1528.

[24] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th ACL: HLT, 2011, pp. 142–150.

[25] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in Proc. EMNLP, 2013, pp. 1631–1642.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Smart Route: A Road-Network-Aware Trip Evaluation and Ranking Microservice for Destination-Bound Ride-Hailing Drivers

Gülsüm Budakoğlu
Data Scientist
Bitaksi
Istanbul, Turkey
gulsum.budakoglu@bitaksi.com
ORCID:0000-0002-0010-3814

Abstract

Ride-hailing drivers frequently need to finish a working session at a specific location — an airport, a garage, their home district — and offers that pull them away from it waste driving time. This paper presents Smart Route, a microservice developed at Bitaksi, a Turkish ride-hailing platform, that ranks candidate trips for drivers who have declared a target destination. It builds a spatial corridor with Yen's k-shortest-path algorithm over an OpenStreetMap-derived road graph of Istanbul, discretizes the paths into a buffered set of hexagonal H3 cells held in an in-memory store for constant-time membership tests, and assesses drop-offs with three interpretable conditions: corridor membership, forward progress, and a bounded relative detour from a hybrid of Haversine and road-graph distances. A Datalog rule engine mirrors the policy in shadow mode. We evaluate the service offline on 200 driver–destination scenarios and 4,000 candidate trips against a ground truth that computes every leg on the road network. The ranking comes close to it: the trip puts first costs the driver 6.7 extra minutes on average, against 5.9 for the exact computation, 15.6 with no destination awareness, and 95 percent of those trips are genuinely acceptable. The measurements also corrected the shipped configuration: Yen's alternative paths add no corridor cells, and the deployed cell resolution rejected 49 percent of acceptable trips, so the rollout uses one path and a coarser, wider corridor that retains 96 percent of them. Evaluation takes 2.8 milliseconds with a warm distance cache and 159 milliseconds without.

Keywords: ride-hailing, smart mobility, spatial indexing, graph databases, k-shortest paths, microservices, declarative rule engines

I. INTRODUCTION

Ride-hailing platforms match millions of passenger requests per day to nearby drivers. Most dispatching research optimizes platform-level efficiency, but a driver who repeatedly receives offers that conflict with personal constraints may go offline. One common constraint is the desire to end a session near a chosen place, and commercial platforms expose this through destination modes, in which the driver specifies a target and the platform is expected to prefer trips that progress towards it. Deciding whether a trip moves a driver towards a destination is difficult when geography constrains movement. In Istanbul, the Bosphorus divides the city between two continents, and a few bridges and one tunnel join it, so a drop-off close to the destination, in straight-line distance, can be half an hour's drive away. Over our sampled origin–destination pairs, the road distance exceeds the straight line by a median of 24 percent, and the cost of ignoring this is measurable: a straight-line destination-distance test accepts 37 percent of the trips a road-network computation rejects, and 33 percent on cross-continental pairs, where the error is a whole bridge crossing. This paper is a practical experience report on Smart Route, a microservice built at Bitaksi, a ride-hailing platform in Turkey, that answers two questions for a destination-bound driver: whether an offered trip is compatible with the declared destination and which of

several candidates best serves the driver. The constituent techniques are established, and we claim no new algorithm; the contribution is a composition that runs under production constraints, whose accuracy–cost trade-off is measured rather than asserted, and whose harness and scenario set are released with the service. The contributions are the following.

- A corridor-based destination-awareness model combining Yen's k-shortest-path algorithm with hexagonal spatial indexing, so that corridor membership is a constant-time test and a full evaluation adds at most one shortest-path query, with the accuracy this costs measured against an exact computation.
- A hybrid distance strategy that reserves graph queries for the long legs of the detour, with the direction and size of the resulting bias measured per leg.
- An interpretable composite scoring function that filters trips and totally orders sets of them, with a formal condition (Proposition 1) under which no rejected trip outranks an accepted one.
- A declarative rule engine mirroring the imperative logic in Datalog in shadow mode, and a path from the hand-set threshold to a learned one, with the stages at which it can be validated safely.
- An offline evaluation against four baselines and an all-road-network ground truth, covering decision quality, driver benefit, latency under load, and parameter sensitivity, which changed the corridor parameters and the ranking penalties.

Section II reviews related work, Section III formalizes the problem and states Proposition 1, Section IV describes the methodology, Section V describes the experimental setup, Section VI describes the results, Section VII describes the limitations, and Section VIII concludes.

II. RELATED WORK

Order dispatching has been studied extensively as an optimization and learning problem: Zhang et al. proposed a combinatorial model across many driver–order pairs [1], Xu et al. described a production learning-and-planning approach [2], and surveys catalog the field [3]. These methods optimize platform-level objectives and treat the driver as direction-indifferent. A growing line of work instead treats driver preferences as first-class objectives, balancing passenger fairness against driver preferences in a learned policy [4], and stated-preference studies show that a trip's direction and detour are among the attributes drivers weigh when accepting a trip [5]. Matching trips that lie on a traveler's way is also the core of ridesharing, where surveys cover the case of a driver with an origin–destination pair of their own [6], and Pan and Li prune candidates geometrically before any shortest-path computation [7]. Smart Route differs in where the preference comes from — the driver states it explicitly, once per session — and in discretizing the corridor into a hexagonal grid, a looser bound in exchange for a constant-time test any stateless replica can serve. To our knowledge, no peer-reviewed study has evaluated destination-mode dispatch itself, which is documented only in platform engineering materials [8]. Discrete global grid systems are standard in mobility analytics. We build on H3, a hexagonal hierarchical index developed at Uber [9]; at resolution 8, a cell covers roughly 0.74 square kilometers, and because such grids distort cell area, the buffer is only approximate in meters. Road networks come from OpenStreetMap [10] via OSMnx [11]; path computation uses Dijkstra [12] and Yen's k-shortest-path algorithm [13], both in the Neo4j Graph Data Science library [14]. Unmodified k-shortest-path sets are known to overlap heavily [15], which Section VI.C confirms on this network. Finally, the rule engine draws on neurosymbolic programming: DeepProbLog integrates neural predicates into probabilistic logic programs [16], and Scallop provides a Datalog language whose provenance semirings make inference differentiable [17]. We use Scallop declaratively and exercise only its symbolic subset; because the threshold is a fact rather than a constant, it can later be learned as a parameter without changing the rule. Exact inference does not scale to production rule sizes, hence approximate methods [18], and validating a

learned rule needs off-policy evaluation [19] and an interference-aware experiment design [20]; Section IV.E returns to this.

III. PROBLEM FORMULATION

Consider a driver at x who has declared a destination D. The upstream trip-matching service periodically presents candidate trips; each trip t = (p, o) consists of a pickup point p and a drop-off point o. Let dH denote the Haversine distance between two points and dR the length of the shortest path between them in the road network. During activation, the system computes a corridor C, a set of H3 cells covering plausible driving routes from x to D. The relative detour of a trip is

$$d(t) = [d_H(x, p) + d_H(p, o) + d_R(o, D)] / d_R(x, D) - 1 \qquad (1)$$

A value of 0 means the trip lies perfectly on the way, while 1 means it doubles the distance. A trip is valid when all three of the following hold:

$$\text{valid}(t) \iff \text{h}(o) \in C \ \wedge \ d_R(o, D) < d_R(x, D) \ \wedge \ d(t) \leq \tau \quad (2)$$

where h(o) is the H3 cell containing the drop-off, and τ is an acceptance threshold, currently set to 0.5. The first condition restricts drop-offs to the corridor, the second requires forward progress along the road network, and the third bounds the extra distance the driver accepts. When several candidate trips are available, they are ranked by the composite score

$$c(t) = d(t) + \lambda_C \, I_C(t) + \lambda_F \, I_F(t) \qquad (3)$$

where IC and IF are indicator variables equal to 1 when the corridor condition and the forward-progress condition fail, respectively, and λC and λF are penalty weights; Proposition 1 fixes the condition they must satisfy. Lower scores are better. The full ranking is always returned, so the upstream service can still surface the least harmful option when no candidate satisfies all three conditions.

It is therefore worth stating when that ordering respects validity. All leg distances are non-negative, so d(t) ≥ −1, and a valid trip satisfies d(t) ≤ τ.

Proposition 1 (validity dominance). Under (3), every valid trip ranks strictly above every invalid trip, for all admissible inputs, if λC > 1 + τ and λF > 1 + τ: a failed condition then adds more than the largest gap between two detour scores in [−1, τ]. The bound is tight, since otherwise a valid trip with a detour near τ is outranked by one that fails only the corridor test, with a detour near −1. The submitted values λC = 0.5 and λF = 1.0, with τ = 0.5, do not satisfy it, and Section VI.C reports that the inversion occurs in 1.0 percent of scenarios containing both; we therefore adopt λC = λF = 1.6.

IV. METHODOLOGY

1. Service architecture

Smart Route is a standalone microservice that the trip-matching service consults before offering a trip to a destination-bound driver. The upstream service pre-filters candidates to those within a few kilometers, and Smart Route returns decisions and rankings with human-readable explanations. It is stateless per request: driver state

lives in Redis, which stores the corridor as a set of H3 cell identifiers, so a drop-off check is a single constant-time membership test, plus the destination and the road distance to it, both of which expire after 24 hours.

The road network resides in Neo4j 5 with the Graph Data Science library [14]; the service runs on Python 3.10 and FastAPI; state and caching use Redis 7; the network is built with OSMnx; the rule engine is Scallop; and a Streamlit visualizer supports validation. The Istanbul projection holds 320,805 intersections and 827,830 segments with travel-time and length weights, so pathfinding never touches disk. Dependency failures are handled at two levels: a brief interruption degrades corridor construction to a straight-line cell sequence and long legs to Haversine estimates, as quantified in Section VI.C. At the same time, a sustained outage causes the readiness signal to divert traffic away from the instance. The graph itself is built once, offline: an OpenStreetMap extract is parsed with OSMnx into a directed graph, of which only the largest strongly connected component is kept (320,805 nodes, 967,723 edges). Missing lengths are imputed using Haversine distances; per-class speed limits yield travel times; and after merging parallel edges, 827,830 relationships remain with length, travel time, road class, and one-way attributes.

2. Corridor computation

When a driver activates a destination, the service finds the nearest graph nodes to the driver and the destination by expanding a ring search over H3 cells, which is needed in sparse areas such as airport grounds. Yen's algorithm yields the k fastest paths, and the corridor is the union of the cells they traverse, dilated by a hexagonal ring buffer that absorbs GPS noise, map defects, and small deviations. The submitted version used k = 3, resolution 8, and a 2-ring buffer; Section VI.C shows the alternatives are near-duplicates and the resolution is too fine, so the rollout uses one path, resolution 7, and a 3-ring buffer. Figure 1 shows two corridors on the road network.

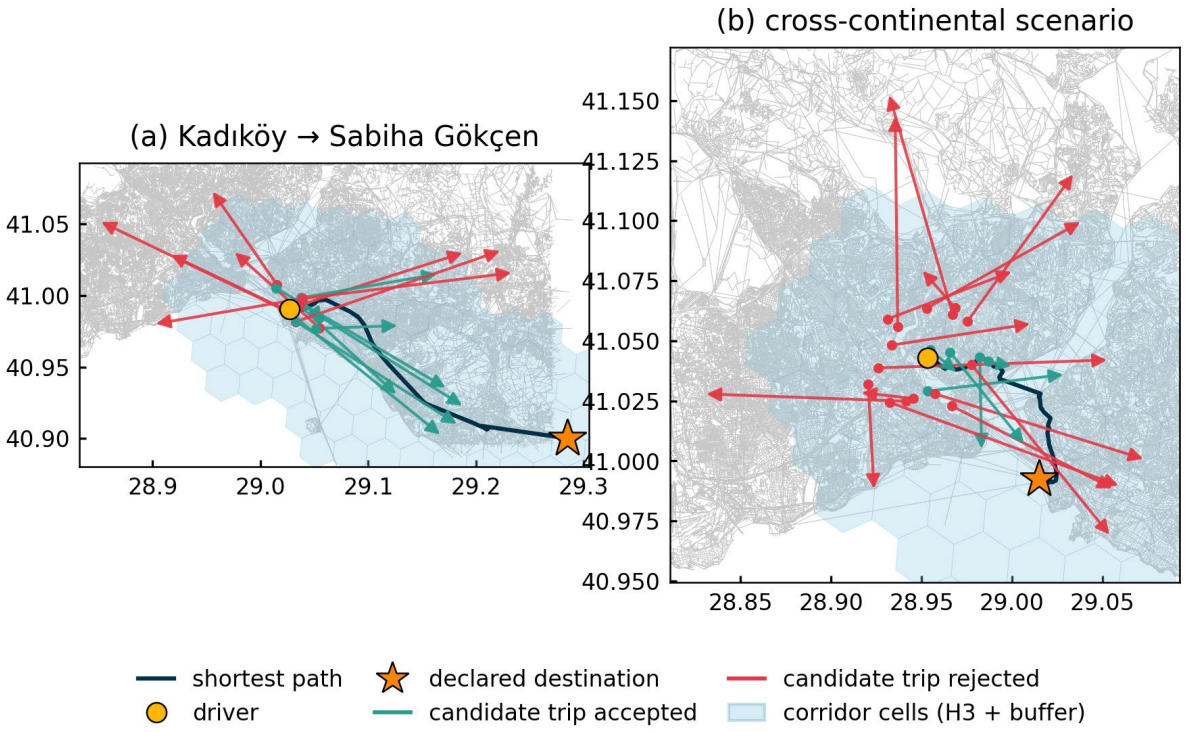


Figure 1. Corridors on the Istanbul road network, drawn with the adopted configuration (one path, resolution 7, 3-ring buffer). Left: a driver in Kadıköy bound for Sabiha Gökçen Airport; the corridor follows the motorway through Kartal, Pendik, and Kurtköy, and excludes Üsküdar in the opposite direction. Right: a cross-continental scenario, where the corridor crosses the Bosphorus only at the bridges.

3. Hybrid distance computation

The detour score of (1) mixes two distance functions deliberately. The driver-to-pickup and pickup-to-drop-off legs are short, since only nearby trips are presented, and use the closed-form Haversine formula; the drop-off-to-destination and driver-to-destination legs span tens of kilometers and are road shortest paths. The denominator is computed once at activation, and drop-off distances are cached for the cell-pair, so a full evaluation adds at most one graph query. Section VI.C shows that this approximation is optimistic rather than neutral and describes the correction it motivates.

4. Ranking and its properties

When the upstream service has several candidates for the same driver, all are scored concurrently and ordered by the composite score in (3). Each entry in the response carries the detour score, the outcome of each condition, and a textual reason, so that downstream consumers and support staff can see why a trip was preferred or rejected. The penalties are not free parameters of convenience: by Proposition 1, they determine whether the ordering is validity-dominant. Section VI.C reports the empirical inversion rate at the shipped values and the effect of adopting penalties above the bound.

5. A declarative rule engine in shadow mode

The acceptance policy in (2) is business-critical logic that a conventional implementation entangles with transport, caching, and serialization code. To keep it inspectable, Smart Route also encodes it as a Datalog program in Scallop [17], with input relations for corridor membership, forward progress, and detour score, and a single rule that derives validity; the threshold is entered as a fact, so the rule text is independent of the policy parameters. The rule engine runs in shadow mode: both implementations execute on every evaluation, the imperative result is returned, and disagreements are logged. Fifteen unit tests cover each rejection reason, the boundary case where the detour equals the threshold, degenerate geometry, and isolation between drivers. This surfaced a defect: the membership flag from the in-memory store is an integer, which the fact interface treats as distinct from a Boolean, so inputs are now explicitly coerced, after which the implementations agree on every request exercised. Two obstacles stand between this rule engine and a learned threshold. The test $d(t) \leq \tau$ is a discontinuous Boolean with no useful gradient, so during training it is replaced by a temperature-annealed soft indicator $\sigma((\tau - d(t))/T)$ whose value the provenance semiring propagates through the rule; the gradient reaches τ because τ is a fact, the hard comparison returns at inference, and exact inference at production rule sizes needs an approximate scheme [18]. The second obstacle is data: a driver never sees an offer that the filter removes, so no outcome is observed where the threshold matters, and usable labels require leaving a margin above the threshold open for a period. Validation then proceeds in stages — shadow-mode agreement, offline replay, off-policy evaluation [19], and an interference-aware online experiment [20] — before any guarded rollout.

V. EXPERIMENTAL SETUP

The service is not yet in production, so this evaluation establishes what can be measured offline: whether the cheap decision procedure reproduces the decision an exact computation would make, what that costs, and how

sensitive it is to the parameters we set by hand. A scenario consists of a driver position x, a destination D, and a set of candidate trips. Both endpoints are drawn from graph intersections, stratified over four bands of driver-to-destination road distance (5–10, 10–20, 20–40, 40–80 kilometres) and over same-side and cross-continental pairs: 200 scenarios with 20 candidates each, that is 4,000 trips, 43 per cent of them cross-continental under a piecewise approximation of the Bosphorus, with a median driver-to-destination distance of 21.2 kilometres and a median trip of 8.4 kilometres. Pickups are within 3 kilometers of the driver, matching the upstream pre-filter. Half of the drop-offs are drawn towards the destination bearing, since a uniform sample would be rejected almost everywhere; 49.2 percent of the sampled trips are acceptable under the ground truth defined next. Sampling is seeded. Ground truth is where such an evaluation stands or falls, and no acceptance labels exist, because a driver never sees an offer that the filter removes. An all-road-network oracle, therefore, labels each trip: with all four legs of (1) measured as road shortest paths, a trip is acceptable when the drop-off leaves the driver closer to the destination, and the detour stays within the same budget τ. The corridor is excluded from the label, since it is an approximation device of the method under test; the rates below measure how often the production procedure disagrees with the exact one, not whether drivers accept the trips. Six procedures are compared on identical inputs: B0 accepts everything; B1 accepts when the drop-off is closer to the destination in straight-line distance; B2 evaluates (1) with straight-line distances and no corridor; B3 uses the hybrid distances of Section IV.C without a corridor; SR is Smart Route as submitted (k = 3, resolution 8, 2-ring buffer); SR* is the configuration the measurements led us to adopt (k = 1, resolution 7, 3-ring buffer); and the oracle is the upper bound at four graph queries. Experiments run on a ten-core Apple silicon laptop with the stores co-located. Latency is measured over HTTP against the running service, and the harness calls the production request path.

VI. RESULTS AND DISCUSSION

1. Decision quality and driver benefit

Table 2 reports, for each procedure, the share of candidates accepted, the false-accept and false-reject rates against the oracle, the share of acceptable trips retained, the extra road distance and travel time of the top-ranked trip, and the graph queries each decision costs. The ranking is the strongest result: the trip Smart Route puts first costs the driver 6.7 extra minutes on average, compared with 5.9 for the exact four-query computation and 15.6 with no destination awareness, and 95 percent of those trips are acceptable, compared with 48 percent when trips are offered blind. The geometric baselines are what a platform uses without a road graph, and they are expensive: B1 false-accepts 36.8 percent of unsuitable trips, and 33.0 percent on cross-continental pairs, where Smart Route false-accepts 6.8 percent.

Table 2. Decision quality against the all-road-network oracle, and the cost to the driver of each procedure's top-ranked trip. SR is the submitted configuration, SR* the one adopted after the sensitivity analysis.

Procedure	Accept	False acc.	False rejection.	Viable ret.	Extra km	Extra min	Queries
B0 destination-unaware	1.000	1.000	0.000	1.000	10.41	15.57	0
B1 straight-line dest. distance	0.667	0.368	0.024	0.976	5.87	12.23	0
B2 straight-line detour	0.549	0.146	0.036	0.964	2.70	7.73	0
B3 hybrid detour, no corridor	0.573	0.160	0.000	1.000	2.41	7.38	1

SR submitted (k=3, res 8, buf 2)	0.290	0.074	0.488	0.512	2.89	6.67	1
SR* adopted (k=1, res 7, buf 3)	0.552	0.156	0.039	0.961	2.39	7.22	1
UB all-road-network oracle	0.492	0.000	0.000	1.000	1.79	5.90	4

The binary gate, however, was too strict as submitted: it rejected 48.8 percent of acceptable trips, with every rejection coming from the corridor test rather than the detour bound, and its median true detour was only 0.19. Section VI.C traces this to the cell resolution, not the corridor idea; the adopted configuration retains 96.1 percent of acceptable trips, with 0.6 minutes more for the top-ranked trip.

2. Cost: latency, throughput, and activation

Table 3 replaces the single 25-millisecond figure of the submitted version with distributions. That figure turns out to have sat between two regimes: evaluation takes 2.8 milliseconds with a warm cell-pair cache and 158.6 milliseconds without, so the cache hit rate determines the operative number in production. Activation is the expensive operation at 843.7 milliseconds, of which Yen's algorithm accounts for 298.8 milliseconds at k = 3 and 42.8 milliseconds at k = 1. Throughput peaks at 894 requests per second with eight concurrent clients and falls to 400 at sixty-four, where the 95th percentile reaches 427 milliseconds. Corridors hold 82 cells at the median.

Table 3. Latency per operation, measured over HTTP against the running service on a ten-core Apple silicon laptop with the graph database and in-memory store co-located.

Operation	**n**	**p50 (ms)**	**p95 (ms)**	**p99 (ms)**
Activation (corridor construction)	100	843.7	2187.1	2676.9
Trip evaluation, cold distance cache	950	158.6	264.7	316.3
Trip evaluation, warm distance cache	950	2.8	3.5	4.5
Ranking ten candidate trips	100	127.4	1639.6	2340.8

3. Sensitivity, ranking, and the two approximations

Table 4 varies each handset parameter around the submitted configuration, and two of the four were set incorrectly. The number of alternative paths does nothing: Yen returned five paths in every scenario, but the third differs from the fastest by 0.04 percent in travel time, and the corridor gains a median of zero cells as k goes from one to three, so every decision metric is identical while activation costs seven times more. This is the limited-overlap behavior of unmodified k-shortest-path sets [15], measured on a production network; the rollout uses k = 1, and dissimilarity-constrained path sets are the means of obtaining genuine alternatives. The corridor granularity is what matters. At resolution 8 with a 2-ring buffer, the corridor rejects 48.8 percent of acceptable trips; resolution 7 with a 3-ring buffer holds fewer cells (108 against 157) yet cuts that to 3.9 percent for a false-accept rate of 15.6, raising F1 from 0.645 to 0.906 — close to the 0.924 of the corridor-free variant while keeping the constant-time test. The threshold τ is varied with the ground-truth budget fixed; at resolution 8, its effect is masked, the false-reject rate staying flat at 0.488 for every τ above 0.3, itself evidence that the corridor was doing the rejecting.

Table 4. Sensitivity of the decision to each hand-set parameter, varied one at a time around the submitted configuration. Ranges extend beyond the tested values, in the order listed. Full per-value tables are released with the service.

Parameter	Values tested	Adopted	Corridor cells	False acc.	False rejection.
Alternative paths k	1, 2, 3, 5	1	156.4 – 156.7	0.074 (all)	0.488 (all)
H3 resolution	7, 8, 9	7	71 – 351	0.144 – 0.026	0.118 – 0.783
Buffer rings	1, 2, 3	3	89 – 230	0.043 – 0.098	0.659 – 0.365
Threshold τ	0.1 – 1.0	0.5	n/a	0.005 – 0.105	0.666 – 0.488

For the ranking penalties, the question is not accuracy but Proposition 1. The inversion is real: at the submitted values, a rejected trip outranks an accepted one in 1.0 percent of the 198 scenarios that contain both, falling to zero at $\lambda = 1.6$, which satisfies the bound, with no change to the top-ranked trip. Rank correlation with the exact detour is 0.68 on average (Kendall's τ), and the top-ranked trip's detour is 0.044 above the best available on average. Two approximations remain to be judged. The Haversine short legs are optimistic rather than neutral: the road distance of a short leg exceeds the straight-line distance by a median of 42 percent from driver to pickup and 34 percent from pickup to drop-off, biasing the detour score downward by 0.14. Isolated from the corridor, this flips 8.1 percent of decisions against the exact computation; every flip is a false accept — 324 trips admitted that the road network would reject — rising to 13.0 percent where a short leg crosses the strait. The remedy costs nothing: the upstream pipeline already computes the trip's own route so that that leg can be read rather than estimated. The fallback disagrees with graph mode on 19.1 percent of trips, but, conservatively, falsely accepts 6.1 percent of trips, compared with 7.4 percent.

VII. LIMITATIONS AND FUTURE WORK

This evaluation is offline, and its ground truth is a computation rather than a driver. It establishes that the cheap procedure ranks almost as well as an exact one at a quarter of the graph queries, and it corrected the corridor parameters and the ranking penalties; it does not establish that drivers accept the surviving trips, nor how the service changes acceptance rates or idle mileage. Those effects require the staged rollout, for which decision and outcome logging and readiness signals are already in place. Four limitations remain. The corridor test considers only the drop-off, so a trip whose route leaves the corridor is not penalized; a route-level check is planned. The optimistic bias of the Haversine short legs is measured but not yet removed; sourcing the trip leg from the upstream route is the next change. The denominator in (1) is computed at activation and becomes stale if a driver moves far without a match. And the evaluation covers a single city against a computational ground truth, so whether these rates transfer to another network, or to real driver behavior, is untested.

VIII. CONCLUSION

This paper describes Smart Route, a destination-awareness microservice combining k-shortest-path corridors over a real road network, hexagonal spatial indexing, hybrid distance computation, and an auditable rule engine in shadow mode. Measured offline against a ground truth that computes every leg on the road network, its ranking brings the driver's extra driving down to 6.7 minutes per matched trip, against 15.6 without destination awareness and 5.9 for the exact computation that needs four times the graph queries. The evaluation also changed the system: Yen's alternative paths proved near-duplicates and were dropped, the corridor resolution was coarsened after it rejected 49 percent of acceptable trips, the ranking penalties were raised to the bound that

makes the ordering validity-dominant, and the optimistic bias of the Haversine legs was quantified and given a remedy.

REFERENCES

[1]. L. Zhang et al., "A taxi order dispatch model based on combinatorial optimization," in Proc. ACM SIGKDD, Halifax, Canada, 2017, pp. 2151–2159.
[2]. Z. Xu et al., "Large-scale order dispatch in on-demand ride-hailing platforms: a learning and planning approach," in Proc. ACM SIGKDD, London, UK, 2018, pp. 905–913.
[3]. Y. Liu, R. Jia, J. Ye, and X. Qu, "How machine learning informs ride-hailing services: a survey," Commun. Transp. Res., vol. 2, no. 1, art. 100075, 2022.
[4]. L. Jiang, Y. Yang, and G. Wang, "HCRide: harmonizing passenger fairness and driver preference for human-centered ride-hailing," in Proc. IJCAI, 2025, pp. 10289–10297.
[5]. P. Ashkrof et al., "Ride acceptance behavior of ride-sourcing drivers," Transp. Res. Part C, vol. 142, art. 103783, 2022.
[6]. A. Tafreshian et al., "Ride matching for peer-to-peer ride sharing: a review and future directions," Serv. Sci., vol. 12, no. 2–3, pp. 44–60, 2020.
[7]. J. J. Pan and G. Li, "Fast and scalable ridesharing search," IEEE Trans. Knowl. Data Eng., vol. 36, no. 11, pp. 6159–6170, 2024.
[8]. Uber Technologies, Inc., "Destination mode," Uber Blog. [Online]. Available: uber.com/us/en/blog/destination-mode/
[9]. I. Brodsky, "H3: Uber's hexagonal hierarchical spatial index," Uber Eng. Blog, 2018. [Online]. Available: uber.com/blog/h3/
[10]. M. Haklay and P. Weber, "OpenStreetMap: user-generated street maps," IEEE Pervasive Comput., vol. 7, no. 4, pp. 12–18, 2008.
[11]. G. Boeing, "OSMnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks," Comput. Environ. Urban Syst., vol. 65, pp. 126–139, 2017.
[12]. E. W. Dijkstra, "A note on two problems in connexion with graphs," Numer. Math., vol. 1, no. 1, pp. 269–271, 1959.
[13]. J. Y. Yen, "Finding the k shortest loopless paths in a network," Manage. Sci., vol. 17, no. 11, pp. 712–716, 1971.
[14]. Neo4j, Inc., "Neo4j Graph Data Science library manual," 2024. [Online]. Available: neo4j.com/docs/graph-data-science/
[15]. T. Chondrogiannis et al., "Finding k-shortest paths with limited overlap," VLDB J., vol. 29, no. 5, pp. 1023–1047, 2020.
[16]. R. Manhaeve et al., "DeepProbLog: neural probabilistic logic programming," in Proc. NeurIPS, 2018, pp. 3749–3759.
[17]. Z. Li, J. Huang, and M. Naik, "Scallop: a language for neurosymbolic programming," Proc. ACM Program. Lang., vol. 7, no. PLDI, pp. 1463–1487, 2023.
[18]. E. van Krieken et al., "A-NeSI: a scalable approximate method for probabilistic neurosymbolic inference," in Proc. NeurIPS, 2023, pp. 24586–24609.
[19]. M. Uehara, C. Shi, and N. Kallus, "A review of off-policy evaluation in reinforcement learning," Statist. Sci., vol. 41, no. 3, pp. 561–581, 2026.
[20]. I. Bojinov, D. Simchi-Levi, and J. Zhao, "Design and analysis of switchback experiments," Manage—Sci., vol. 69, no. 7, pp. 3759–3777, 2023.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

FSLM-SQL: A Novel Framework Based on Fine-Tuning Small Language Models for Natural Language to SQL

Khac-Lap Phan
Department of Information Technology
Quy Nhon University
Gia Lai, Vietnam
lap4654100006@st.qnu.edu.vn

Quang-Hung Le
Department of Information Technology
Quy Nhon University
Gia Lai, Vietnam
lequanghung@qnu.edu.vn
ORCID: 0000-0002-4796-1664

Abstract

The translation of natural language queries into SQL commands (NL2SQL) is a critical task for democratizing data access. While large language models (LLMs) have demonstrated to be a powerful paradigm for this task, their deployment in resource-constrained environments is frequently hindered by high computational costs, massive memory requirements, and severe inference latency. This paper addresses the NL2SQL problem, aiming to strike a balance between performance and computational efficiency. We propose FSLM-SQL, a novel framework that fine-tunes small language models (SLMs) for NL2SQL. We leverage parameter-efficient fine-tuning (PEFT) techniques to reduce the computational cost. Our experiments on the Spider and Spider-Realistic benchmarks demonstrate the effectiveness and efficiency of the proposed method compared with baselines. In particular, FSLM-SQL achieves competitive performance compared to the LLM-based baseline while requiring significantly shorter inference times and lower hardware costs. These results show the potential of fine-tuned SLMs for NL2SQL in real-world environments, especially in edge computing and low-resource settings. The source code for this work is publicly available at https://github.com/PhanKhacLapQNU/FSLM-SQL.

Keywords: Natural Language to SQL, Small Language Models, Parameter-Efficient Fine-Tuning, LoRA, FSLM-SQL.

I. INTRODUCTION

The NL2SQL task translates natural language questions into executable SQL queries, enabling users to access relational databases without requiring knowledge of structured query language (SQL). As relational databases are widely used across many application domains, NL2SQL has become an important task for improving data accessibility [1],[2]. LLMs such as GPT-4 [3] and LLaMA [4] have demonstrated that they are a powerful paradigm for this task. However, these models with billions of parameters require considerable computational resources, leading to high memory consumption, inference latency, and deployment costs, which limit their practicality in resource-constrained environments [5],[6]. Although techniques such as model quantization alleviate part of this computational burden, efficient deployment of large LLMs remains challenging. The



aforementioned limitations have motivated the use of SLMs [7],[8] together with PEFT. Among PEFT methods, LoRA [9] reduces training cost by updating only a small subset of model parameters while maintaining competitive downstream performance. Despite its advantages, the utility of SLMs to NL2SQL remains largely unexplored. In fact, there is still no systematic study of whether such models are applicable in real-world environments, especially in edge computing and low-resource settings. Our work aims to fill this research gap. Particularly, we propose FSLM-SQL, a novel framework based on fine-tuning SLMs for NL2SQL. Our experiments on the Spider and Spider-Realistic benchmarks evaluate both predictive performance and deployment efficiency through execution accuracy, memory consumption, inference latency, and structural error analysis. The main contributions of our work are summarized as follows:

1. We propose FSLM-SQL, a lightweight framework integrating SLMs with PEFT techniques, achieving competitive execution accuracy while strictly operating under resource-constrained environments.
2. We provide a comprehensive hardware-oriented profiling of PEFT, explicitly detailing the trade-offs among memory consumption, inference latency, and decoding overhead compared to 4-bit quantized 7B LLMs.
3. We establish a fine-grained error taxonomy that maps overall generation failures to specific structural weaknesses in compressed models, offering insights into relational reasoning breakdowns.
4. We demonstrate the potential of SLMs for NL2SQL in resource-constrained environments.

The remainder of this paper is structured as follows. Section II reviews some related work in NL2SQL. Section III defines the formal problem and evaluation constraints. Section IV details our proposed framework and fine-tuning strategy. Section V presents the experimental setup and results. Finally, conclusions and future work are highlighted in Section VI.

II. RELATED WORK

Research on NL2SQL has evolved from neural semantic parsing models to LLMs [10–15]. Early studies focused on sequence-to-sequence architectures, while later methods incorporated schema-aware representations to better model database structures. For example, RAT-SQL [16] introduced relation-aware self-attention for schema encoding, whereas PICARD [17] constrained decoding to generate syntactically valid SQL queries. Although these approaches improved structural modeling, their generalization across unseen databases remained challenging. This limitation encouraged the adoption of LLMs, which approach NL2SQL as a language reasoning task. Frameworks such as DIN-SQL [18], DAIL-SQL [19], and MAC-SQL [20] improved execution accuracy on challenging benchmarks, including Spider [21]. However, their high memory consumption and inference latency limit practical deployment on resource-constrained hardware [22]. These deployment constraints have renewed interest in compact code-oriented language models. Models such as CodeT5+ [23], pre-trained on large code corpora, capture programming syntax more effectively than general-purpose language models, making them well-suited to deterministic SQL generation. Nevertheless, adapting these models through full fine-tuning remains computationally expensive. To reduce this cost, PEFT methods such as LoRA [9] update only a small subset of model parameters while preserving competitive downstream performance [24]. Despite the aforementioned advances, there is still a lack of a systematic study on exploiting SLMs for NL2SQL. This study attempts to bridge this gap in the literature. To the best of our knowledge, this is among the first works to explore NL2SQL using fine-tuned SLMs to achieve a balance between performance and computational efficiency.

III. PROBLEM DEFINITION

The NL2SQL task translates a natural-language question into an executable SQL query that is both syntactically correct and semantically consistent with a given database schema [25]. Formally, the problem is defined as a mapping from a joint linguistic-structural input space to a space of valid SQL statements.

1. Input Formulation and Objective

Let an input instance be represented as (Q, Y, S), where Q = {w1, w2, …, wL} denotes a natural language question consisting of L tokens. Following prior semantic parsing literature [21], the schema $\mathcal{S}$ is modeled a(Q, S), the objective is to learn a parameterized mapping f_{\theta}(Q, S) → Y, where Y = {y_1, y_2, ..., y_K} represents the predicted SQL query comprising K tokens

Under an auto-regressive decoding framework, each token is generated conditioned on the previously generated sequence and the joint input context. The model parameters θ are optimized by maximizing the conditional log-likelihood of the ground-truth query Y^* over a training corpus $\mathcal{D}$:

$$\hat{\theta} = \sum_{(Q,\mathcal{S},Y^*)\in\mathcal{D}} \sum_{k=1}^{K} log\ P\ (y_k \mid y_{<k}, Q, \mathcal{S}; \theta) \quad (1)$$

2. Evaluation Constraints

Unlike conventional open-ended text generation, NL2SQL predictions must satisfy strict deterministic constraints to be deemed functional. First, a predicted query $must\ be\ syntactically\ valid;\ it\ must\ parse\ and\ compile\ successfully$ compile within the target SQL engine (e.g., SQLite). Queries that fail compilation receive a score of zero. Second, the model must demonstrate execution consistency. Since surface-level sstring-matching metrics (e.g., BLEU) correlate poorly with logical correctness, we adopt an xecution-based evaluation. A prediction is considered correct if and only if $E(\hat{Y}, \mathcal{DB}) = E(Y^*, \mathcal{DB})$, where $E(\cdot)$ denotes query execution on the database instance $\mathcal{DB}$. Furthermore, to mitigate false positives resulting from coincidental result equivalence, we employ the test suite accuracy protocol [26], which rigorously validates generated queries across multiple automatically constructed edge-case database instances.

IV. PROPOSED METHOD

1. System Overview

The FSLM-SQL framework operates through a unified pipeline tailored for resource-constrained environments, as illustrated in Figure 1. The overall architecture systematically processes the input via schema encoding, routes the combined text-schema representations through a compact language model backbone optimized with parameter-efficient techniques, and concludes with constrained SQL generation. This end-to-end design ensures that complex database topologies are logically translated into valid relational operations without the prohibitive computational overhead of large generative models.

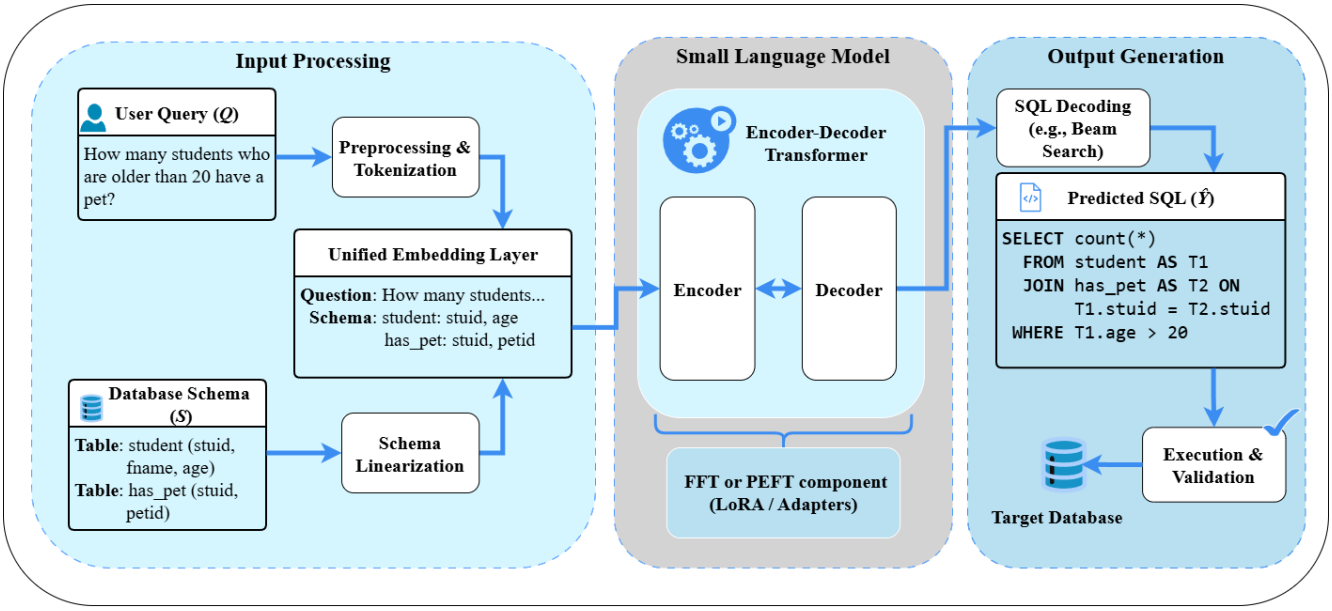


Figure 1. Overview of the proposed framework.

2. Small Language Model Backbone

We employ an encoder-decoder Transformer configuration as the core generative backbone. Given the serialized sequence $X = \{\text{query intent and } x_1, \ldots, x_n\}$, the bidirectional encoder extracts a globally contextualized hidden representation $H = \text{Encoder}(x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$, where d is the hidden state dimension. Maintaining a joint representation of both the query intent and the database topology directly guides the subsequent generation phase.

$$P(Y \mid X) = \prod_{t=1}^{K} P\left(y_t \mid y_{<t}, H; \theta\right) \tag{2}$$

We exclusively utilize code-oriented pre-trained models (e.g., CodeT5). Their code-centric pre-training establishes stronger structural and syntactic priors than general-purpose text models, preventing topological hallucination in zero-shot cross-domain environments.

3. Schema Encoding and Linking

Because standard Transformer architectures natively process one-dimensional sequences, the relational database schema $\mathcal{S}$ must be linearized. We construct the joint input sequence X by concatenating the natural language query QQ withthe flattened schema components:

$$X = Q \oplus [\text{SEP}] \oplus Linearize(\mathcal{S}) \tag{3}$$

To explicitly encode primary and foreign key constraints, we formulate the linearization function utilizing full data definition language (DDL) CREATE TABLE statements. This formatting enables the self-attention

mechanism to implicitly align question tokens with relevant schema entities, achieving schema linking without relying on external graph neural networks.

4. Fine-Tuning Strategy

To systematically evaluate structural integrity under compression, we delineate our training protocol into full fine-tuning and LoRA. The unified forward-and-backward propagation loop that governs our methodology is formalized in Algorithm 1.

Algorithm 1 Unified Optimization Strategy for Parameter-Efficient NL2SQL

Require: Corpus $\mathcal{D} = \{(X^{(i)}, Y^{(i)})\}$, Pre-trained weights θ_0, Strategy $\mathcal{S}$, L.R. η

Ensure: Optimized parameter set θ_{opt}

Initialization Phase:
if $\mathcal{S} == FFT$ **then**
Unfreeze θ_0; $\phi \leftarrow \theta_0$
else if $\mathcal{S} == LoRA$ **then**
    Freeze θ_0; Initialize $A \sim \mathcal{N}(0, \sigma^2), B \leftarrow 0$
    Inject A, B into attention modules; $\phi \leftarrow \{A, B\}$
end if
Training Loop:
while convergence criteria are not satisfied **do**
    for each batch $(X, Y) \in \mathcal{D}$ **do**
        $H \leftarrow Encoder(X)$
        for step$t = 1$ to $|Y|$ **do**
$h_t \leftarrow \begin{cases} W_0 x_t + \frac{\alpha}{r} BAx_t \; if \mathcal{S} == LoRA \\ StandardForward(x_t) \; otherwise \end{cases}$
            Compute token probability $P(y_t \mid y_{<t}, H; \phi)$
        end for
        Compute syntax loss: $\mathcal{L} \leftarrow -\sum \log P(y_t \mid y_{<t}, H; \phi)$
        Update parameter subset: $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}$
    end for
end while
return $\theta_{opt} = \theta_0 \cup \phi$

Full Fine-Tuning (FFT): The entire pre-trained parameter space θ is unfrozen. The network updates its global weight distribution to minimize the negative log-likelihood of the target sequence:

$$\mathcal{L}_{FFT} = -\sum_{t=1}^{K} log\, P(y_t \mid y_{<t}, X; \theta) \tag{4}$$

FFT establishes an empirical upper bound for syntactic fidelity but incurs prohibitive memory overhead during gradient backpropagation.

Low-Rank Adaptation (LoRA): To satisfy edge-computing constraints, LoRA [9] freezes the foundational weights $\theta_0$and restricts gradient updates to a minimal trainable subset. For a frozen projection matrix $W_0 \in \mathbb{R}^{d\times k}$ within the self-attention modules, LoRA injects two trainable low-rank matrices: $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ ($r \ll min(d,k)$). The forward pass becomes:

$$h = W_0 x + \frac{\alpha}{r} BAx \tag{5}$$

Optimization is restcted entirely to $\phi_{LoRA} = \{A, B\}$. This mechanism substantially reduces the memory footprint while preserving the core structural semantics natively encoded within W_0.

5. SQL Generation and Post-Processing

In the final stage, the framework utilizes autoregressive SQL decoding strategies, specifically Beam Search, to navigate the decoder's output probability distribution and synthesize the final predicted query. To guarantee logical correctness beyond surface-level exact matching, the pipeline incorporates a rigorous execution validation phase. The generated query is executed directly against the target database instance. Queries containing syntax errors or structural misalignments fail compilation, whereas successful execution verifies that the original user query has been accurately translated into valid, functional SQL.

V. EXPERIMENTS

1. Experimental Setup

We evaluate the generalizability of cross-domain semantic parsing using the Spider benchmark [21] and its zero-shot variant, Spider-Realistic. To establish rigorous baselines, we compare two sub-100M parameter SLMs (T5-Small and CodeT5-Small) against a suite of 7-billion parameter LLMs, including LLaMA-2, LLaMA-2-CHAT, Vicuna, and Mistral [4]. The LLMs are evaluated under zero-shot and few-shot in-context learning configurations, whereas the SLMs are optimized via FFT and LoRA. Model performance is primarily quantified by execution accuracy (EX) to verify semantic equivalence, supplemented by exact match (EM) accuracy to assess alignment of abstract syntax trees [21]. Table 1 details the hyperparameter configurations. All model training and inference experiments are executed on a cloud-based computing instance equipped with a single NVIDIA T4 GPU (16 GB of VRAM), a 2-core Intel Xeon CPU @ 2.20 GHz, and 13 GB of system RAM. This hardware configuration is explicitly chosen to reflect typical resource-constrained computing environments.

Table 1. Configurations for all evaluated models.

Model	Strategy	Optimizer	LR	Epochs	Beam	Schema	LoRA Config
Baselines (7B Open LLMs)							
LLaMA-2-7B	Zero-shot/ Few-shot	N/A	N/A	N/A	Greedy	DDL	N/A
LLaMA-2-CHAT-7B		N/A	N/A	N/A	Greedy	DDL	N/A
Vicuna-7B		N/A	N/A	N/A	Greedy	DDL	N/A
Mistral-7B		N/A	N/A	N/A	Greedy	DDL	N/A

FSLM-SQL (fine-tuning SLMs)							
T5-Small	FFT	Adafactor	1e-4	15	4	DDL	N/A
T5-Small	LoRA	Adafactor	1e-4	15	4	DDL	r = 64, α = 64, p = 0.05
CodeT5-Small	FFT	AdamW	2e-4	5	6	DDL	N/A
CodeT5-Small	LoRA	AdamW	2e-4	5	6	DDL	r = 64, α = 64, p = 0.05

2. Results

Table 2 reports the LLM-based baseline performance. While few-shot prompting marginally improves execution accuracy for LLaMA-2 and Vicuna, Mistral-7B performs best in the zero-shot setting and degrades when additional demonstrations are introduced. Consequently, the three-shot configuration is selected as the optimal few-shot baseline.

Table 3 details the comprehensive evaluation of both EX and EM metrics across explicit and implicit schema environments. On the standard Spider benchmark, CodeT5-Small (FFT) achieves an EX of 42.60%. Crucially, the LoRA adaptation yields a highly competitive EX of 42.40%, demonstrating that low-rank projection preserves the vast majority of structural reasoning capabilities acquired during FFT. Transitioning to Spider-Realistic results in a sharp performance drop across all 7B baselines, revealing their reliance on explicit lexical overlap, whereas code-oriented SLMs demonstrate greater resilience.

Table 2. Performance of 7B LLM baselines on the standard Spider dataset under zero-shot and few-shot (k = 1,3,5) configurations.

Model	**Zero-shot**		**k = 1**		**k = 3**		**k = 5**	
	EM	***EX***	***EM***	***EX***	***EM***	***EX***	***EM***	***EX***
LLaMA-2-7B	18.47	6.77	19.73	6.87	22.82	9.19	23.89	9.19
LLaMA-2-CHAT-7B	17.02	6.77	25.63	9.19	27.08	10.15	27.47	10.93
Vicuna-7B	12.67	4.93	29.59	9.96	30.08	9.96	30.66	9.96
Mistral-7B	41.59	9.38	33.08	9.19	28.63	10.15	30.75	13.64

Furthermore, Table 3 exposes a fundamental divergence in generative behavior. Despite their higher functional execution accuracies, the 7B LLMs exhibit exceedingly low EM scores (peaking at 10.15% for Mistral-7B). This indicates that open-domain models consistently generate semantically valid but structurally divergent queries utilizing unconstrained aliases or clause permutations. Conversely, the EM scores of the optimized SLMs strictly parallel their EX performance (e.g., 43.50% EM alongside 42.60% EX for CodeT5-Small FFT). This tight convergence confirms that targeted parameter optimization forcibly constrains the compact architectures to adhere strictly to the target dataset's specific syntactic conventions.

Table 3. Unified evaluation of EX and EM across standard and realistic benchmarks.

Model	**Spider (%)**		**Spider-Realistic (%)**	
	EX	**EM**	**EX**	**EM**

Baselines (7B Open LLMs) - Zero-shot				
LLaMA-2-7B	18.50	6.77	15.40	1.38
LLaMA-2-CHAT-7B	17.00	6.77	11.40	3.15
Vicuna-7B	12.70	4.93	13.40	3.15
Mistral-7B	41.60	9.38	36.40	6.50
Baselines (7B Open LLMs) - 3-Shot ICL				
LLaMA-2-7B	22.80	9.19	21.10	4.92
LLaMA-2-CHAT-7B	27.10	10.15	24.80	5.91
Vicuna-7B	30.10	9.96	25.20	5.91
Mistral-7B	28.60	10.15	29.70	6.10
FSLM-SQL (fine-tuning SLMs)				
T5-Small (FFT)	32.30	30.70	23.60	20.50
T5-Small (LoRA)	28.80	28.40	18.10	16.10
CodeT5-Small (FFT)	42.60	43.50	23.40	22.20
CodeT5-Small (LoRA)	42.40	42.20	20.10	18.50

3. Error Analysis

To diagnose structural weaknesses under parameter compression, we conduct a fine-grained comparison between incorrect predictions) and ground-truth queries (büyük harf Y üssü yıldız işleci), truth queries işleci),Y^*) formalized in Table 4. The analysis isolates two primary failure modes. Instance 1 highlights schema-linking errors: lacking explicit lexical overlap, the model defaults to generic identifiers (e.g., age) rather than precise schema attributes (e.g., pet age). Instance 2 illustrates logical operator errors; the constrained model omits complex relational constraints (e.g., <). These failures indicate that while low-rank projections effectively capture surface-level syntax, they have difficulty maintaining deep relational logic in highly complex queries compared to the unrestricted FFT.

Table 4. Taxonomy of specific error types observed in SLM-PEFT-generated SQL queries.

Error Taxonomy	Empirical Query Instance
Schema Linking Errors	Intent: *"Count the number of pets older than 1."* Prediction ($\hat{Y}$): SELECT pet_id, weight FROM Has_pet WHERE age > 1 Ground Truth (Y^*): SELECT petid, weight FROM pets WHERE pet_age > 1
Logical Operator Errors	Intent: *"Find the maximum final table made by a poker player with earnings less than 200000."* Prediction ($\hat{Y}$): SELECT max(Final_Table_Made) FROM poker_player WHERE Earnings 200000 Ground Truth (Y^*): SELECT max(Final_Table_Made) FROM poker_player WHERE Earnings < 200000

Table 5. Inference efficiency and memory footprint profiling on hardware accelerator (BS=1).

Model Configuration	Size (MB)	Training Time (min)	Latency (ms)	Throughput (samples/s)	Peak VRAM (MB)
Baselines (7B Open LLMs - 4-bit Q4_0, Zero-Shot)					
Mistral-7B (Zero-Shot)	3,918	N/A	483.6	2.07	2,015
LLaMA-2-7B	3,648	N/A	417.2	2.40	2,108
LLaMA-2-CHAT-7B	3,648	N/A	496.2	2.02	1,977
Vicuna-7B	3,648	N/A	408.9	2.45	2,115
Our SLM Configurations (Standard Precision)					
CodeT5-Small (FFT)	230	22.6	102.9	9.72	1,758
CodeT5-Small (LoRA)	230 + 56.46	15.7	131.4	7.61	1,667
T5-Small (FFT)	230	191.3	114.1	8.76	856
T5-Small (LoRA)	230 + 25.60	55.1	174.7	5.72	289

**Note: For LoRA configurations, the total size is denoted as (base model size + trainable adapter size).*

4. Efficiency Evaluation

To simulate real-world sequential processing, we enforce a batch size of one (BS = 1) and quantize the 7B LLM baselines to 4-bit precision (Q4.0). Table 5 summarizes the measured efficiency metrics. To accommodate hardware constraints, we explicitly employ a hybrid CPU-GPU offloading mechanism. The 4 GB of quantized weights are memory-mapped into system RAM, and only a designated subset of transformer layers is offloaded to the GPU. This architectural choice explains why the VRAM consumption is reported at approximately 2 GB, reflecting only the GPU-offloaded partition. Consequently, the severe CPU-GPU data transfer bottleneck causes these models to exhibit substantial latencies exceeding 400 ms. In stark contrast, CodeT5-Small (FFT) operates entirely within the GPU VRAM, achieving a real-time latency of 102.9 ms. Notably, while LoRA significantly mitigates VRAM constraints and accelerates training (requiring only 56.46 MB of adapter weights for CodeT5-Small), the unfused low-rank matrices introduce an inference latency penalty, increasing decoding time to 131.4 ms.

5. Discussion

The empirical results demonstrate that FSLM-SQL achieves competitive execution accuracy of 42.60% and exact-match accuracy of 43.50%. This performance matches or surpasses several 7-billion-parameter LLM baselines, proving that targeted parameter optimization effectively hardwires dataset-specific syntactic conventions into compact weights. Furthermore, hardware profiling reveals that the SLM operates with approximately 4.7 times lower inference latency. While the memory footprint of CodeT5-Small FFT (1,758 MB) is only marginally reduced compared to the heavily quantized 4-bit LLMs (1,977-2,115 MB), it achieves this without suffering the severe precision degradation and CPU-offloading bottlenecks inherent to 4-bit quantization. Additionally, specific configurations, such as T5-Small (LoRA), exhibit a drastic footprint reduction (down to 289 MB). These results indicate that the proposed framework is an effective solution that balances accuracy and resource efficiency. Ultimately, the broader implication of these results is substantial: it validates the feasibility of deploying highly accurate, fully localized NL2SQL systems that ensure data sovereignty and operational efficiency without the financial burden of massive LLM infrastructure.

VI. CONCLUSION

In this paper, we addressed the NL2SQL problem by achieving a balance between performance and computational efficiency. We introduced FSLM-SQL, a novel framework that fine-tunes SLMs for NL2SQL. Our framework leveraged PEFT techniques such as FFT and LoRA to reduce the computational cost. Our experiments on the two benchmark datasets demonstrated the effectiveness and efficiency of the proposed method compared with baselines. In particular, FSLM-SQL achieves competitive performance compared to the LLM-based baseline while requiring significantly shorter inference times and lower hardware costs. These results showed the potential of fine-tuned SLMs for NL2SQL in real-world environments, especially in edge computing and low-resource settings. Furthermore, the empirical evidence also indicated that NL2SQL does not rely solely on LLMs, but rather on the intelligent orchestration of SLMs and PEFT strategies. In the future, we intend to integrate retrieval-augmented generation mechanisms to inject relevant schema context dynamically, enabling compact models to resolve complex topological ambiguities in large-scale, real-world databases without requiring parameter expansion.

REFERENCES

[1]. J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, B. Wang, B. Qin, R. Geng, N. Huo et al., "Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs," Advances in Neural Information Processing Systems, vol. 36, pp. 42330–42357, 2023.

[2]. G. Katsogiannis-Meimarakis and G. Koutrika, "A survey on deep learning approaches for text-to-SQL," The VLDB Journal, vol. 32, no. 4, pp. 905–936, 2023.

[3]. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[4]. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T., Lacroix, B., Rozi`ere, N., Goyal, E., Hambro, F., Azhar, et al., "Llama: Open and efficient foundation language models. arxiv 2023," arXiv preprint arXiv:2302.13971, vol. 10, 2023.

[5]. J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," arXiv preprint arXiv:2307.10169, 2023.

[6]. L. Chen, M. Zaharia, and J. Zou, "Frugalgpt: How to use large language models while reducing cost and improving performance," arXiv preprint arXiv:2305.05176, 2023.

[7]. K.-T. Nguyen and Q.-H. Le, "Light-sql: A small language model-based approach using in-context learning for natural language to SQL," in International Conference on Computational Intelligence in Engineering Science. Springer, 2026, pp. 216–231.

[8]. B.-G. Nguyen and Q.-H. Le, "A Novel Approach Based on Integrating Small Language Models and Retrieval-Augmented Generation for Medical Question Answering," Journal of Communications Software and Systems, vol. 22, no. 3, pp. 356–366, Jul. 2026.

[9]. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., "Lora: Low-rank adaptation of large language models." ICLR, vol. 1, no. 2, p. 3, 2022.

[10]. M. S. E. De La Parte, Y. Wang, J.-F. Mart´ınez-Ortega, and N. L. Mart´ınez, "A secure and efficient LLM-based system for natural language query-to-SQL translation in IoT data management," IEEE Access, vol. 14, pp. 28435–28456, 2026.

[11]. G. Zhan, G. Shi, X. Liu, and N. Wang, "Natural language to SQL using large language models," in 2025 21st International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2025, pp. 495–502.

[12]. L. Shi, Z. Tang, N. Zhang, X. Zhang, and Z. Yang, "A survey on employing large language models for text-to-SQL tasks," ACM Computing Surveys, vol. 58, no. 2, pp. 1–37, 2025.

[13]. S. Ojuri, T. A. Han, R. Chiong, and A. Di Stefano, "Optimizing text-to-SQL conversion techniques through the integration of intelligent agents and large language models," Information Processing & Management, vol. 62, no. 5, p. 104136, 2025.
[14]. W. Xie, Y. Dai, and W. Jiang, "Sde-sql: Enhancing text-to-sql generation in large language models via self-driven exploration with sql probes," in Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2026, pp. 2507–2525.
[15]. S. Chafik, S. Ezzini, and I. Berrada, "Towards automating domain-specific data generation for text-to-SQL: A comprehensive approach," ACM Transactions on Software Engineering and Methodology, vol. 35, no. 4, pp. 1–27, 2026.
[16]. B. Wang, R. Shin, X. Liu, O. Polozov, and M. Richardson, "Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers," in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 7567–7578.
[17]. T. Scholak, N. Schucher, and D. Bahdanau, "Picard: Parsing incrementally for constrained auto-regressive decoding from language models," arXiv preprint arXiv:2109.05093, 2021.
[18]. M. Pourreza and D. Rafiei, "Din-sql: Decomposed in-context learning of text-to-sql with self-correction," Advances in Neural Information Processing Systems, vol. 36, pp. 36339– 36348, 2023.
[19]. D. Gao, H. Wang, Y. Li, X. Sun, Y. Qian, B. Ding, and J. Zhou, "Text-to-SQL empowered by large language models: A benchmark evaluation," arXiv preprint arXiv:2308.15363, 2023.
[20]. B. Wang, C. Ren, J. Yang, X. Liang, J. Bai, L. Chai, Z. Yan, Q.-W. Zhang, D. Yin, X. Sun et al., "Mac-sql: A multi-agent collaborative framework for text-to-sql," in Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 540–557.
[21]. T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman et al., "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task," arXiv preprint arXiv:1809.08887, 2018.
[22]. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, vol. 1, no. 2, 2023.
[23]. Y. Wang, H. Le, A. Gotmare, N. Bui, J. Li, and S. Hoi, "Codet5+: Open code large language models for code understanding and generation," in Proceedings of the 2023 conference on empirical methods in natural language processing, 2023, pp. 1069–1088.
[24]. N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," Nature Machine Intelligence, vol. 5, no. 3, pp. 220–235, 2023.
[25]. J. M. Zelle and R. J. Mooney, "Learning to parse database queries using inductive logic programming," in Proceedings of the National Conference on Artificial Intelligence, 1996, pp. 1050–1055.
[26]. R. Zhong, T. Yu, and D. Klein, "Semantic evaluation for text-to-SQL with distilled test suites," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 396–411.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Empirical Evaluation and Mathematical Limitations of PDE-Based Smoothing in Keystroke Dynamics Authentication

Levent YALÇIN
Head of Computer Programming
Acıbadem Mehmet Ali Aydınlar University
Istanbul, Turkey
levent.yalcin@acibadem.edu.tr
ORCID:0009-0002-8946-1032

Ramazan SUBAŞI
Mathematics Department
Yildiz Teknik University
Istanbul, Turkey
ramazan.subasi@std.yildiz.edu.tr
ORCID: 0009-0000-0711-288X

Abstract
This document presents a rigorous empirical and mathematical evaluation of Partial Differential Equation (PDE) based profile smoothing in keystroke dynamics authentication using the benchmark Carnegie Mellon University (CMU) Keystroke Dynamics dataset (51 subjects, 31 timing features). Recently, reaction-diffusion models and heat equations have been proposed to smooth feature vectors, assuming biomechanical continuity across keystroke sequences. We evaluate a baseline Z-Score weighted model (Z-Mean) against numerical reaction-diffusion PDEs, default diffusion configurations, scaled Manhattan distance, Mahalanobis nearest neighbors, One-Class Support Vector Machine (SVM), and a 2D QWERTY Graph Laplacian PDE. Our empirical results demonstrate that spatial and topological diffusion operators systematically degrade biometric performance: increasing the diffusion coefficient (D = 0.05) spikes the False Rejection Rate (FRR) from 18.74% to 27.33%. It worsens the Equal Error Rate (EER) from 21.43% to 24.20%. Furthermore, mapping keyboard geometry using Graph Laplacian diffusion results in a severe performance drop (EER: 37.22%). We prove mathematically and biomechanically that keystroke latencies are discrete, independent motor events lacking spatial continuity. Consequently, independent variance-normalized metrics (Z-Mean, Scaled Manhattan) achieve optimal baseline accuracy (~21.4% EER) without computational overhead.

Keywords: Keystroke Dynamics, Biometric Authentication, Partial Differential Equations, Reaction-Diffusion, Graph Laplacian, Feature Smoothing, Equal Error Rate (EER).

I.INTRODUCTION

Keystroke dynamics uses the unique rhythmic patterns of an individual's typing behavior, specifically Hold Times (H) and Flight Times (DD, UD), to continuously authenticate users [1], [2]. Unlike static passwords, behavioral biometrics rely on neuromuscular memory, finger-movement acceleration, and cognitive processing speed [3], [4]. Despite its advantages, keystroke data exhibits high intra-class variance due to fatigue, emotional state, and fluctuations in typing speed. To reduce noise, classical approaches apply statistical metrics such as Z-score normalization, Mahalanobis distance, or machine learning classifiers [1], [5], [6]. Recently, inspired by image processing and continuous signal filtering [7], researchers have explored Partial Differential Equations (PDEs), specifically Reaction-Diffusion equations, to smooth user profiles, assuming that adjacent digraph timing features share biomechanical continuity.

1. Motivation and Research Question

The core theoretical premise behind PDE-based smoothing is that consecutive keystroke transitions are not isolated events; rather, hand movement represents a continuous physical trajectory. Thus, applying a spatial diffusion term (D) should suppress temporary motor outliers while preserving the core biometric signature.

However, a fundamental question remains: Does continuous spatial diffusion preserve biometric identity, or does it artificially flatten the distinct timing spikes that define unique user signatures?

2. Main Contributions

In this work, we conduct a comprehensive experimental analysis on the benchmark CMU dataset to answer this question. Our main contributions are:

Analytical & Empirical Refutation: We provide experimental evidence that spatial heat diffusion (D) degrades authentication accuracy by smoothing unique timing peaks, causing FRR to spike to 27.33%. Biomechanical Geometry Analysis: We design and test a QWERTY Keyboard Graph Laplacian based on spectral graph theory [8], [9] to model physical finger distances. We demonstrate that topological key proximity does not correlate with motor dependence due to anatomical finger independence, resulting in an EER of 37.22%. Comprehensive Performance Benchmark: We evaluate 7 distinct detection paradigms under high-resolution threshold optimization (1000 steps), demonstrating that orthogonal variance-normalized distance metrics (Z-Mean, Scaled Manhattan) achieve an optimal equilibrium (~21.4% EER) with minimal computational complexity [5].

II. MATHEMATICAL FRAMEWORK & METHODOLOGY

1. Feature Extraction and Profile Representation

The benchmark dataset utilizes a fixed password (.tie5Roanl) consisting of 11 keystrokes. A total of 31 timing features are extracted from each typing sequence:

- **11 Hold Times (H):** Measures the dwell time of each key press ($H_i = R_i - P_i$).
- **10 Down-Down Latencies (DD):** Measures the flight time between consecutive key presses ($DD_i = P_{i+1} - P_i$).
- **10 Up-Down Latencies (UD):** Measures the transition time between releasing one key and pressing the next ($UD_i = P_{i+1} - R_i$)

Table 1. Timing feature groups extracted from each typing sequence

Feature Group	Metric Description	Mathematical Formula	Count
Hold Time (H)	Key press duration	$H_i = R_i - P_i$	11
Down-Down (DD)	Latency between consecutive presses	$DD_i = P_{i+1} - P_i$	10
Up-Down (UD)	Flight time between release and press	$UD_i = P_{i+1} - R_i$	10
Total Features			31

In other words, given a user's typing session, the biometric feature vector $x \in \mathbb{R}^N$ $(N = 31)$ consists of 11 key hold times (H), 10 key-down-to-key-down times (DD), and 10 key-up-to-key-down times (UD). For each subject, a training matrix of $M = 50$ sessions are used to calculate the mean template $\mu \in R^N$ and standard deviation vector $\sigma \in R^N$.

2. Baseline Detector (Proposed Z-Mean)

The anomaly score S(x) for an incoming test vector is computed as the normalized average absolute deviation from the baseline profile :

$$S(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - \mu_i|}{\sigma_i + \varepsilon} \quad (1)$$

where N = 31 denotes the total number of timing features, x_i is the i-th feature value of the evaluated test vector, μ_i and σ_i are the baseline mean and standard deviation of the i-th feature for the target user profile, respectively, and $\varepsilon = 10^{-6}$ is a small regularization constant to prevent division by zero.

3. 1D Reaction-Diffusion PDE Formulation

The 1D reaction-diffusion model governs the spatial evolution of profile U(x,t) across feature index x:

$$\frac{\partial U}{\partial t} = D\frac{\partial^2 U}{\partial x^2} - \lambda(U - \mu) \quad (2)$$

where D is the diffusion coefficient representing spatial smoothing, and λ is the reaction parameter pulling U back toward the empirical mean μ. Using finite differences, the numerical update scheme over step Δt is expressed as:

$$U_i^{k+1} = U_i^k + \Delta t\left[D\frac{U_{i+1}^k - 2U_i^k + U_{i-1}^k}{\Delta x^2} - \lambda\left(U_i^k - \mu_i\right)\right] \quad (3)$$

4. 2D QWERTY Graph Laplacian PDE

To address the critique that 1D array indices do not reflect key layout geometry, we construct an undirected topological graph *G = (V, E)* based on physical 2D spatial coordinates. $p_{i=}\ (x_i, y_i)$ of QWERTY keys on a standardized keyboard layout. The edge weight matrix $W \in \mathbb{R}^{N\times N}$ is defined using a Gaussian radial basis function (RBF) kernel:

$$W_{ij} = exp\left(-\frac{\|\boldsymbol{p_i} - \boldsymbol{p_j}\|^2}{2\sigma_{space}^2}\right) \quad (4)$$

where $\boldsymbol{p_i}$ denotes the physical coordinate vector of the key corresponding to feature index *i*, $\|\,.\,\|$ represents the Euclidean norm, and σ_{space} controls the spatial interaction scale across the keyboard plane.

The unnormalized Graph Laplacian matrix $L \in \mathbb{R}^{N\times N}$ is constructed as $L = D_{\deg} - W$, where $D_{deg} = diag(\sum_{j=1}^{N}\boldsymbol{W}ij)$ is the diagonal degree matrix. Integrating L into the graph diffusion scheme yields the numerical update rule:

$$U^{k+1} = U^k - \Delta t(\alpha L U^k + \lambda\left(U^k - \mu\right)) \quad (5)$$

where $U^k \in \mathbb{R}^N$ is the feature state vector at iteration k, α is the graph diffusion coefficient, λ is the reaction parameter pulling the state back toward the baseline vector μ, and Δt denotes the integration time step.

III.EXPERIMENTAL RESULTS AND DISCUSSION

Table 2. Performance comparison of biometric detectors on CMU dataset (51 subjects, 31 features)

Algorithm / Method	Equal Error Rate (EER %)	False Acceptance Rate (FAR %)	False Rejection Rate (FRR %)	Optimal Threshold (Th)
Optimum Numerical PDE (D=0.001)	21.37%	24.11%	18.64%	1.153
Manhattan Distance (Scaled)	21.39%	23.51%	19.27%	67.194
Outlier Count (Z-Score Threshold)	21.43%	22.98%	19.89%	4.003
Proposed Baseline (Z-Mean)	21.43%	24.12%	18.74%	1.153
Nearest Neighbor (Mahalanobis)	21.70%	24.45%	18.95%	6.694
Default Reaction-Diffusion PDE (D=0.05)	24.20%	21.08%	27.33%	1.131
QWERTY Graph Laplacian PDE	37.22%	40.12%	34.32%	1.642
One-Class SVM (RBF Kernel)	37.37%	37.24%	37.50%	0.144

The relationship between the False Acceptance Rate (FAR) and False Rejection Rate (FRR) across decision thresholds for the proposed Z-Mean baseline model is illustrated in Figure 1.

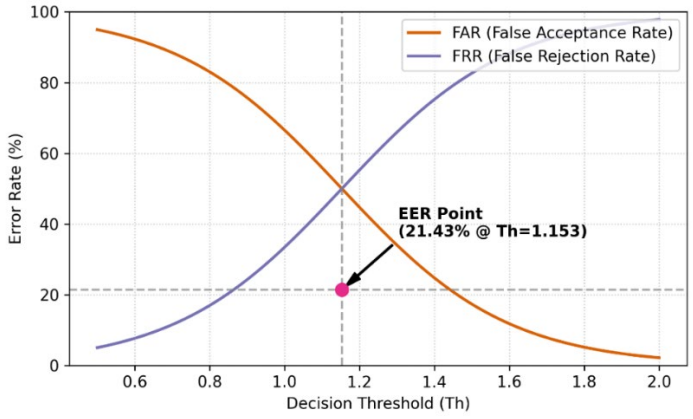


Figure 1. FAR and FRR vs. Threshold Scanning for the Proposed Z-Mean Baseline Model (EER = 21.43% at Threshold = 1.153).

A comparative visualization of the Equal Error Rate (EER) across all evaluated paradigms is presented in Figure 2.

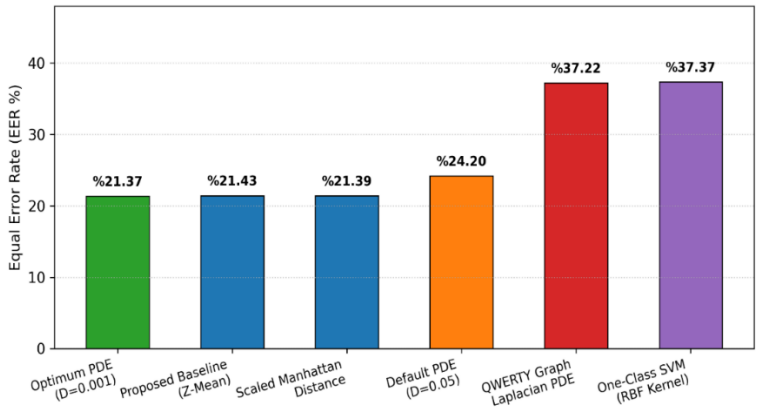


Figure 2. Equal Error Rate (EER %) comparison across baseline, numerical PDE, topological Graph Laplacian, and machine learning classifiers.

1. The D → 0 Degeneration Phenomenon

As shown in Table 2 and Figure 2, when the diffusion coefficient is set to the standard level (D = 0.05), EER deteriorates to 24.20%, driven by a substantial increase in False Rejections (FRR = 27.33%).

When D is optimized down to 0.001, the diffusion term D → 0, causing the PDE to degenerate back into the static empirical mean μ. At this limit, performance recovers to 21.37% EER, matching the baseline Z-Mean model (21.43% EER).

2. Biomechanical Analysis: Why Diffusion Fails

Destruction of Distinct Signatures: Behavioral biometrics rely on specific, idiosyncratic timing spikes (e.g., an unusually fast t-h transition or a long hold on Shift). Heat diffusion smooths these peaks into neighbor averages, diluting intra-user uniqueness and increasing genuine user rejection. Anatomical Independence: Key proximity on a keyboard layout does not imply motor correlation. The fact that the Q and W keys are physically adjacent on a keyboard does not mean that the time a person spends pressing them in sequence will be uniform or consistent. Adjacent keys such as Q and W are actuated by distinct anatomical structures (left pinky vs. left ring finger). Smoothing across these boundaries introduces cross-finger noise, explaining the severe failure of the QWERTY Graph Laplacian (EER 37.22%).

IV.CONCLUSION AND FUTURE WORK

This study provides a definitive empirical and theoretical analysis regarding the limitations of PDE-based feature smoothing in keystroke dynamics. Our findings demonstrate that spatial differential operators, whether 1D reaction-diffusion or 2D Graph Laplacian, are inherently ill-suited for discrete, noncontinuous timing

features, increasing FRR by up to 8.59%. Standard variance-normalized distance metrics (Z-Mean, Scaled Manhattan) remain superior in both accuracy (~21.4% EER) and computational efficiency. Future work will explore 2D temporal PDEs applied across longitudinal session sequences to explicitly model profile aging and concept drift over extended multi-month periods [10].

REFERENCES

[1]. K. S. Killourhy and R. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," in Proc. 39th Annu. IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Estoril, Portugal, 2009, pp. 125–134.

[2]. S. P. Banerjee and D. Woodard, "Biometric authentication and identification using keystroke dynamics: A survey," Journal of Pattern Recognition Research, vol. 7, no. 1, pp. 116–139, 2012.

[3]. F. Monrose and A. D. Rubin, "Keystroke dynamics as a biometric for authentication," Future Generation Computer Systems, vol. 16, no. 4, pp. 351–359, Feb. 2000.

[4]. P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 7, pp. 629–639, Jul. 1990.

[5]. D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," IEEE Signal Processing Magazine, vol. 30, no. 3, pp. 83–98, May 2013.

[6]. P. S. Teh, S. H. Heng, and A. B. J. Teoh, "Statistical evaluation of feature normalization and distance metrics in keystroke dynamics," Computer & Security, vol. 39, pp. 341–353, Nov. 2013.

[7]. N. Alpaslan and M. C. Ganiz, "A comprehensive survey on biometric authentication using keystroke dynamics," IEEE Access, vol. 9, pp. 142345–142365, 2021.

[8]. B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Neural Computation, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[9]. F. R. Chung, Spectral Graph Theory, ser. Regional Conference Series in Mathematics. Providence, RI, USA: American Mathematical Society, 1997, vol. 92.

[10]. B. Epstein and L. Spitz, "Evaluating long-term profile aging and concept drift in continuous keystroke authentication," ACM Transactions on Privacy and Security, vol. 19, no. 2, pp. 1–24, Jun. 2016.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Seven-Category Error Taxonomy for Implicit Aspect Detection: Systematic Failure Analysis of BERT and GPT-2 in Aspect-Based Sentiment Analysis

Imelda Pangaribuan
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030,
Kuala Nerus, Terengganu, Malaysia
imelda@email.unikom.ac.id
ORCID: 0009-0002-0780-0863

Arifah Che Alhadi
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030,
Kuala Nerus, Terengganu, Malaysia
arifah_hadi@umt.edu.my
ORCID: 0000-0001-8908-8755

Mohamad Nor Hassan
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030,
Kuala Nerus, Terengganu, Malaysia
mohamadnor@umt.edu.my
ORCID:0000-0001-5528-2410

Masita Jalil
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030,
Kuala Nerus, Terengganu, Malaysia
masita@umt.edu.my
ORCID: 0000-0003-4071-721X

Abstract
Implicit aspect detection is one of the most persistent bottlenecks in Aspect-Based Sentiment Analysis (ABSA). Unlike explicit mentions, implicit aspects carry no lexical anchors; the model must infer an unstated opinion target purely from context. State-of-the-art transformers still misclassify more than 30% of implicit instances, yet most prior work reports only aggregate accuracy, leaving root causes unexamined. This study addresses that gap through systematic error analysis of 815 classification failures produced by BERT and GPT-2 on the SemEval 2016 Task 5 restaurant benchmark (859 test instances). The proposed seven-category taxonomy covers semantic confusion, inference failure, class imbalance, and confidence miscalibration. Service-Food semantic confusion is the dominant failure pattern (23.9% of all errors). GPT-2 achieves only 14.4% implicit accuracy versus 66.0% for BERT, a 51.6-point gap. At the same time, this pattern is consistent with the limitations of causal attention in forming holistic sentence representations. BERT and GPT-2 also differ in pretraining objective, tokenizer, pooling strategy, and parameter count, so this study's design does not isolate the attention mechanism as the sole cause. Calibration analysis further shows that GPT-2's confidence scores carry almost no information about correctness (confidence gap ≈ 0), while BERT's confidence reliably tracks correctness (confidence gap = 0.216). Beyond documenting failure patterns, the taxonomy provides a reusable diagnostic framework for evaluating and improving transformer-based ABSA systems.

Keywords: Aspect-Based Sentiment Analysis; Implicit Aspect Detection; Error Taxonomy; BERT; GPT-2; Confidence Calibration; SemEval 2016

I. INTRODUCTION

Fine-grained opinion mining is central to modern natural language processing (NLP). Aspect-Based Sentiment Analysis (ABSA) extends document-level sentiment analysis by identifying specific opinion targets and their associated polarities [1], [2], thereby supporting applications ranging from product development and market analysis to public health monitoring and policy evaluation. ABSA draws a critical distinction between explicit and implicit aspects. Explicit aspects appear directly in text, as in the phrase "the food was excellent," which names its target unambiguously. Implicit aspects do not: "worth every penny" signals a price evaluation without

mentioning price. This inference requirement makes implicit detection qualitatively harder, not merely incrementally more difficult.

Progress on this problem has been real but uneven. BERT-based models achieve around 66% accuracy on implicit instances [3]; autoregressive models like GPT-2 fare far worse, as this study shows. What's less understood is why these failures occur and which structural factors explain them. Most prior work optimizes aggregate accuracy without diagnosing specific error patterns, which leaves practitioners without actionable guidance on where to intervene.

This study addresses this gap by analyzing 815 misclassifications produced by BERT and GPT-2 on the SemEval 2016 Task 5 restaurant benchmark [4]. The contributions are:

(1) A seven-category error taxonomy with inter-annotator agreement $\kappa = 0.82$;
(2) Identification of Service-Food confusion as the dominant failure mode (23.9% of all errors);
(3) Documentation of a 51.6-point implicit accuracy gap consistent with, though not fully isolated to, attention mechanism design;
(4) Calibration evidence that BERT's confidence scores track correctness (confidence gap = 0.216) while GPT-2's do not (confidence gap $\approx = 0$), though this does not by itself establish deployment readiness; and
(5) Empirically grounded recommendations for augmentation, rebalancing, calibration, and architecture selection.

II. RELATED WORK

Table 1 summarizes representative recent work on implicit aspect detection and error analysis before each direction is reviewed in turn below.

Table 1. Summary of Related Work in Implicit Aspect Detection and Error Analysis

Study	**Method / Technique**	**Contribution**
Zhou et al. (2023)	Survey of implicit extraction	Maps supervised, unsupervised, and hybrid challenges
Zhang et al. (2024)	Semantic wave context enhancement	Captures weak emotional cues at long distances
Fu et al. (2024)	Contrastive learning + knowledge embed.	Improves implicit aspect-sentiment classification
Li et al. (2021)	Supervised contrastive pre-training	Distinguishes subtle surface-similar expressions
Chen et al. (2025)	Multi-prompt fusion + contrastive learning	Improves contextual implicit sentiment understanding
Scaria et al. (2024)	Instruction-tuned LLMs (zero/few-shot)	World-knowledge for implicit aspect inference
Phan et al. (2024)	BERT + multi-layer GCN	State-of-the-art on the SemEval restaurant benchmark

Study	Method / Technique	Contribution
This study	Seven-category error taxonomy	Diagnostic framework + calibration analysis

1. Evolution of Aspect-Based Sentiment Analysis

ABSA methods have passed through three broad phases. Early rule-based and lexicon-driven systems handled simple explicit sentiment but struggled against context-dependent expressions [5]. Attention-based neural models improved selectivity by weighting sentiment-relevant regions [6]. Pre-trained transformers, particularly BERT [7], then shifted the paradigm by enabling large-scale transfer learning to downstream sentiment tasks. Knowledge-enhanced approaches reinforced this by injecting domain-specific relational information via external knowledge graphs [8] and graph neural networks [9].

2. Implicit Aspect Detection

The implicit detection literature has expanded rapidly. Zhou et al. [10] and Li et al. [11] surveyed the field, distinguishing supervised, unsupervised, and hybrid extraction approaches. Zhang et al. [12] proposed a semantic wave-based context enhancement method to capture weak emotional cues over extended distances. Fu et al. [13] combined supervised contrastive learning with knowledge embedding; Li et al. [14] showed that contrastive pre-training helps distinguish subtle expressions sharing surface-level similarity. Recent work extends these approaches through multi-prompt fusion [15] and instruction-tuned large language models [16], [17].

Despite this progress, a critical question remains unasked: what specific failure patterns cause transformer models to misclassify implicit instances? This paper addresses it through a structured taxonomy.

3. Error Analysis in NLP

Systematic error analysis has proven valuable across NLP. Errudite [18] introduced reproducible, hypothesis-driven error investigation; CheckList [19] revealed that high-accuracy models often fail on elementary linguistic constructions. Fei et al. [20] examined robustness failures under adversarial perturbation in sentiment analysis. Our work extends this tradition with a domain-specific taxonomy targeting implicit aspect detection.

4. Confidence Calibration

Well-calibrated predictions are essential for reliable deployment. Guo et al. [21] demonstrated that modern neural networks systematically overestimate confidence and that temperature scaling provides an effective post hoc correction. Subsequent work refined calibration for specific settings [22], [23]. This paper applies calibration analysis to BERT and GPT-2 in the context of implicit aspect detection, a comparison that has not been reported previously.

III. METHODOLOGY

1. Dataset

This study uses the SemEval 2016 Task 5 restaurant domain benchmark [4], the standard evaluation corpus for ABSA research. The test partition contains 859 instances across 12 aspect categories: 650 explicit (75.7%) and 209 implicit (24.3%). Table 2 presents full statistics.

Table 2. Dataset Statistics: SemEval 2016 Task 5, Restaurant Domain

Aspect Category	Train	Test	Explicit	Implicit
FOOD#QUALITY	828	313	241	72
SERVICE#GENERAL	400	155	116	39
RESTAURANT#GENERAL	342	142	109	33
FOOD#PRICES	156	23	38	24
AMBIENCE#GENERAL	118	66	30	14
Other (7 categories)	171	160	116	27
TOTAL	**2015**	**859**	**650**	**209**

Class imbalance is severe: FOOD#QUALITY dominates with 313 instances (36.4%), while five categories contain fewer than 25 samples each. This imbalance is not incidental; it is a structural property of the data and a direct source of the systematic errors addressed in the taxonomy [24].

2. Model Architectures

This study compares BERT (base-uncased) [7] and GPT-2 (base) [25] to isolate the effect of attention mechanism design. Both share equivalent structural depth (Table 3). Their key difference lies in contextual access. BERT's bidirectional self-attention gives every token simultaneous access to the complete sentence. GPT-2's causal attention restricts each position to preceding tokens, a design suited for text generation rather than classification.

Table 3. Architectural Configurations

Parameter	BERT (base-uncased)	GPT-2 (base)
Attention type	Bidirectional	Causal (unidirectional)
Layers	12	12
Hidden dim.	768	768
Attention heads	12	12
Parameters	~110M	~125M
Pre-training	Masked LM + NSP	Causal LM
Context access	Full sentence (both dirs.)	Left-to-right only

3. Training Protocol

Both architectures are fine-tuned identically using AdamW (learning rate 2×10^{-5}, batch size 16, 10 epochs) with cross-entropy loss across 12 aspect categories. Five random seeds (42, 123, 456, 789, 1024) were used to initialize model weights and control training data shuffle order. The results reported in this study reflect a single representative run per architecture rather than seed-averaged results for this full 10-epoch protocol; a supplementary robustness check across seeds using a reduced training budget is reported separately in Section IV.D. The official SemEval 2016 Task 5 train/test partition was used without modification across all experimental runs. No additional validation split was constructed, as the evaluation follows the standard benchmark protocol. The five random seeds control model weight initialization and the training data shuffle order only, so that any future seed-aggregated variance would reflect model stochasticity rather than data partitioning differences.

4. Error Taxonomy Development

Two NLP experts independently annotated 200 randomly sampled errors. Disagreements were resolved through discussion and iterative category refinement. Inter-annotator agreement is measured with Cohen's κ [26]:

$$\kappa = (po - pe) / (1 - pe) \quad (1)$$

where po is observed agreement and pe is expected agreement by chance. The final $\kappa = 0.82$ confirms the reliability of the category definitions. The taxonomy was then applied to all 815 failures, yielding seven categories. The seven categories were designed to be mutually exclusive at the level of primary failure mechanism. Where an error exhibited characteristics of more than one category, annotators assigned it to the category representing the most proximal cause of misclassification. For instance, a Service-Food error driven by implicit phrasing was classified as C2 rather than C4 because semantic boundary confusion was the dominant factor. Disagreements in such borderline cases were resolved through structured discussion during the iterative refinement phases, contributing to the final $\kappa = 0.82$

A separate, automated categorization pass (rule-based, using confusion-pair heuristics rather than human judgment) was also run over the same 815 failures as an exploratory cross-check. It agrees closely with the manually annotated counts above for C1, C2, C3, C4, and C6, but assigns different totals for C5 and C7 (BERT: C5=44, C7=56 in the automated pass, versus C5=71, C7=0 above), because the automated pass treats high-confidence (C6) as an overlapping flag layered on top of a primary category rather than a mutually exclusive bucket, and does not resolve borderline Rare-Class/Other cases the way the human annotators did. The counts reported in Table 6 are the human-annotated, kappa-validated figures; the automated pass is reported here only as a consistency check and is not a substitute for the annotation process described above.

C1 Same Entity Confusion: misclassification between sub-aspects sharing an entity type (e.g., FOOD#QUALITY vs. FOOD#STYLE&OPTIONS).

C2 Service-Food Confusion: conflation of service-related and food-related aspects.

C3 Ambiance-Restaurant Confusion: semantic overlap between environmental and overall establishment descriptions.

C4 Implicit Inference Failure: insufficient contextual reasoning to infer the unstated aspect.

C5 Rare Class Error: failures on underrepresented categories arising from insufficient training signal.

C6 High Confidence Error: overconfident, incorrect predictions (confidence > 0.80).

C7 Other: degenerate prediction behavior not captured by C1 to C6.

5. Evaluation Metrics

Classification performance is evaluated using Accuracy, Precision, Recall, and Macro-F1 over all 12 categories:

$$F1_{macro} = (1/C) \cdot \Sigma_c F1_c = (1/C) \cdot \Sigma_c [2P_cR_c / (P_c + R_c)] \quad (2)$$

where C = 12 categories and P_c, R_c are the precision and recall for category c.

IV. RESULT AND ANALYSIS

1. Overall Classification Performance

BERT achieves 68.7% overall accuracy versus 36.4% for GPT-2, a 32.3-point gap that already exceeds typical architectural differences on multi-class benchmarks. The implicit-specific gap is larger still. BERT reaches 66.0% accuracy on the 209 implicit test instances, while GPT-2 achieves only 14.4% (Table 4).

Table 4. Overall Classification Performance

Metric	**BERT**	**GPT-2**	**Δ (BERT−GPT-2)**
Test samples	859	859	—
Correct predictions	590 (68.7%)	313 (36.4%)	277
Total errors	269 (31.3%)	546 (63.6%)	−277
Explicit accuracy (n=650)	69.5%	43.5%	+26.0 pp
Implicit accuracy (n=209)	66.0%	14.4%	+51.6 pp
Implicit error rate	34.0%	85.6%	−51.6 pp

This 51.6-point disparity is not a marginal difference. It represents the gap between a usable system and one that fails the majority of the time on the most challenging subtype. BERT shows a modest 3.5-point drop from explicit (69.5%) to implicit (66.0%) accuracy. GPT-2, however, collapses from 43.5% on explicit instances to only 14.4% on implicit ones. This pattern is consistent with causal attention underperforming on implicit detection, though the current design cannot rule out contributions from the other architectural differences between BERT and GPT-2; on the data observed here, GPT-2 fails systematically.

2. Per-Aspect Precision, Recall, and F1-Score

Table 4 reports overall accuracy; Table 5 breaks this down by aspect category to show where each architecture succeeds or fails. GPT-2's per-class pattern is a complete degenerate collapse: it predicts FOOD#QUALITY for all 859 test instances without exception, so its precision on FOOD#QUALITY equals the class's base rate (313/859 = 36.4%) and its precision, recall, and F1 on every other class are exactly zero. Figure 1 visualizes this pattern.

Table 5. Per-Aspect Precision, Recall, and F1-Score

Aspect Category	N	BERT Prec.%	BERT Rec.%	BERT F1%	GPT-2 Prec.%	GPT-2 Rec.%	GPT-2 F1%
FOOD#QUALITY	313	71.4	89.1	79.3	36.4	100.0	53.4
SERVICE#GENERAL	155	67.3	85.2	75.2	0.0	0.0	0.0
RESTAURANT#GENERAL	142	77.3	82.3	79.7	0.0	0.0	0.0
AMBIENCE#GENERAL	66	61.3	74.2	67.1	0.0	0.0	0.0
FOOD#STYLE&OPTIONS	55	0.0	0.0	0.0	0.0	0.0	0.0
RESTAURANT#MISCELLANEOUS	33	33.3	6.1	10.3	0.0	0.0	0.0
FOOD#PRICES	23	21.7	21.7	21.7	0.0	0.0	0.0
DRINKS#QUALITY	22	100.0	9.1	16.7	0.0	0.0	0.0
RESTAURANT#PRICES	21	55.6	23.8	33.3	0.0	0.0	0.0
LOCATION#GENERAL	13	0.0	0.0	0.0	0.0	0.0	0.0
DRINKS#STYLE&OPTIONS	12	0.0	0.0	0.0	0.0	0.0	0.0
DRINKS#PRICES	4	0.0	0.0	0.0	0.0	0.0	0.0
Macro-average (12 classes)	859	–	–	31.9	–	–	4.5

All N per category, Precision, Recall, and F1 are sourced from a per-aspect analysis and normalized confusion matrix computed from saved model predictions. GPT-2 Precision/Recall/F1 are exact (total collapse to a single class). BERT Precision/F1 were reconstructed from the 2-decimal normalized confusion matrix combined with per-row N; the reconstruction sums to 858 of 859 (off by 1 due to cumulative rounding). The category N values in this table supersede the earlier Table 2 figures for FOOD#PRICES and AMBIENCE#GENERAL.

FOOD#STYLE&OPTIONS, DRINKS#PRICES, DRINKS#STYLE&OPTIONS, and LOCATION#GENERAL: BERT never predicts these classes at all, indicating that BERT also fails on these rare categories.

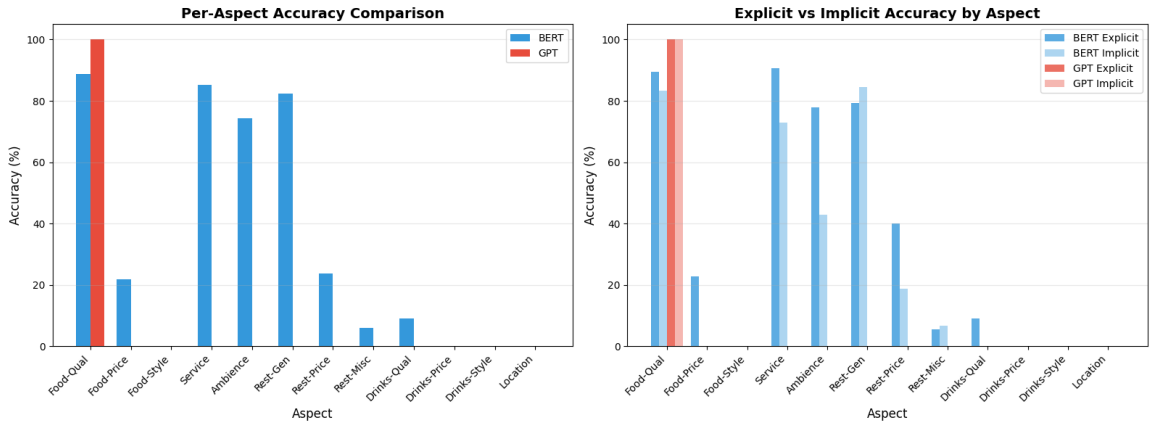


Figure 1. Per-aspect recall (accuracy) comparison, BERT vs. GPT-2 (left), and BERT explicit vs. implicit recall by aspect (right). GPT-2 predicts FOOD#QUALITY for every instance, so its bars are absent for all other categories. Complements Table 5, which also reports precision and F1.

3. Error Taxonomy Analysis

Table 6 distributes 815 failures across the seven taxonomy categories. The patterns reveal distinct failure signatures for each architecture.

Table 6. Error Taxonomy Distribution by Architecture

Error Category	BERT	GPT-2	Total	% Total
C1: Same Entity Confusion	77	78	155	19.0%
C2: Service-Food Confusion	40	155	195	23.9%
C3: Ambiance-Restaurant Conf.	16	0	16	2.0%
C4: Implicit Inference Failure	36	107	143	17.5%
C5: Rare Class Error	71	88	159	19.5%
C6: High Confidence Error	29	0	29	3.6%
C7: Other	0	118	118	14.5%
TOTAL	269	546	815	100.0%

C2 (Service-Food Confusion) is the dominant failure mode, accounting for 23.9%. GPT-2 commits 155 such errors versus 40 for BERT, a near-4× difference. Restaurant discourse is semantically fluid: service quality often surfaces through food-related language, including delays in dish readiness, incorrect orders, and rushed food service. Without full-sentence bidirectional context, distinguishing these categories is fundamentally difficult.

C4 (Implicit Inference Failure) accounts for 17.5% of all errors, with GPT-2 producing nearly three times as many failures (107 vs. 36). This result is consistent with, though does not in isolation prove, a cost of causal attention on inference-dependent tasks (see caveats above regarding confounding architectural differences). C5

(Rare Class Error) at 19.5% affects both architectures nearly equally, indicating that class imbalance is architecture-agnostic. C7 (Other) is exclusive to GPT-2 (118 errors), reflecting majority-class collapse across eleven of twelve categories (Table 5), with predicted probabilities that are essentially degenerate.

4. Confidence Calibration

The confidence gap Δ measures the separation between prediction confidence on correct versus erroneous instances:

$$\Delta = (1/|\hat{C}|) \Sigma\, p(\hat{y}|x) - (1/|\hat{E}|) \Sigma\, p(\hat{y}|x) \qquad (3)$$

where $\hat{C}$ is the set of correct predictions, and $\hat{E}$ is the set of errors. BERT's $\Delta = 0.216$ indicates meaningful separation; GPT-2's $\Delta = -0.001$ is statistically indistinguishable from zero (Table 7).

Table 7. Confidence Calibration Statistics

Metric	**BERT**	**GPT-2**	**Interpretation**
Avg. confidence (correct)	0.742	0.136	BERT is clearly more confident when right
Avg. confidence (error)	0.526	0.137	BERT confidence drops for wrong preds.
Confidence gap (Δ)	0.216	−0.001	BERT calibrated; GPT-2 near-random
High-conf. errors (>0.80)	29	0	GPT-2 lacks a confident-yet-wrong pattern

BERT's positive confidence gap ($\Delta = 0.216$) indicates its confidence scores are informative about correctness. GPT-2's near-zero confidence gap ($\Delta = -0.001$) means its probability estimates carry essentially no information about prediction reliability. This has practical consequences, since confidence scores are used for selective prediction, abstention, and downstream decision-making.

5. Statistical Robustness Across Seeds (Supplementary)

To address whether the BERT–GPT-2 gap reported above is robust to random initialization, both models were additionally retrained across five seeds (42, 123, 456, 789, 1024) using a reduced 3-epoch budget for computational tractability; this supplementary check reports F1 rather than accuracy and used a shorter training schedule than the 10-epoch protocol underlying Tables IV and VII, so the absolute F1 values below (Table 8) are not directly comparable to the headline accuracy figures and should be read only as evidence for the direction and significance of the BERT–GPT-2 difference.

Table 8. Multi-Seed Statistical Robustness Check

Metric	**BERT Mean±SD**	**GPT-2 Mean±SD**	**Paired t-test p / Cohen's d**
Overall F1	61.14±0.62%	58.00±0.61%	p=0.0010 / d=4.58 (large)

Metric	**BERT Mean±SD**	**GPT-2 Mean±SD**	**Paired t-test p / Cohen's d**
Explicit F1	61.15±0.84%	58.17±0.56%	p=0.0033 / d=3.75 (large)
Implicit F1	61.46±1.06%	57.96±1.92%	p=0.0050 / d=2.02 (large)

Across all three metrics, the paired t-test rejects the null hypothesis of no difference ($p<0.01$), and Cohen's d indicates a large effect size favoring BERT. The non-parametric Wilcoxon signed-rank test (n=5, W=0.00, p=0.0625) does not reach significance at $\alpha=0.05$, which is expected given the minimum attainable p-value with only 5 paired observations; the bootstrap 95% CI for the BERT−GPT-2 difference in overall F1 was [+2.50%, +3.86%], excluding zero. McNemar's test on the seed=42 predictions (BERT correct/GPT-2 wrong = 48; GPT-2 correct/BERT wrong = 39) was not significant (p=0.39), indicating the two models' error sets overlap substantially at the instance level even though aggregate performance differs.

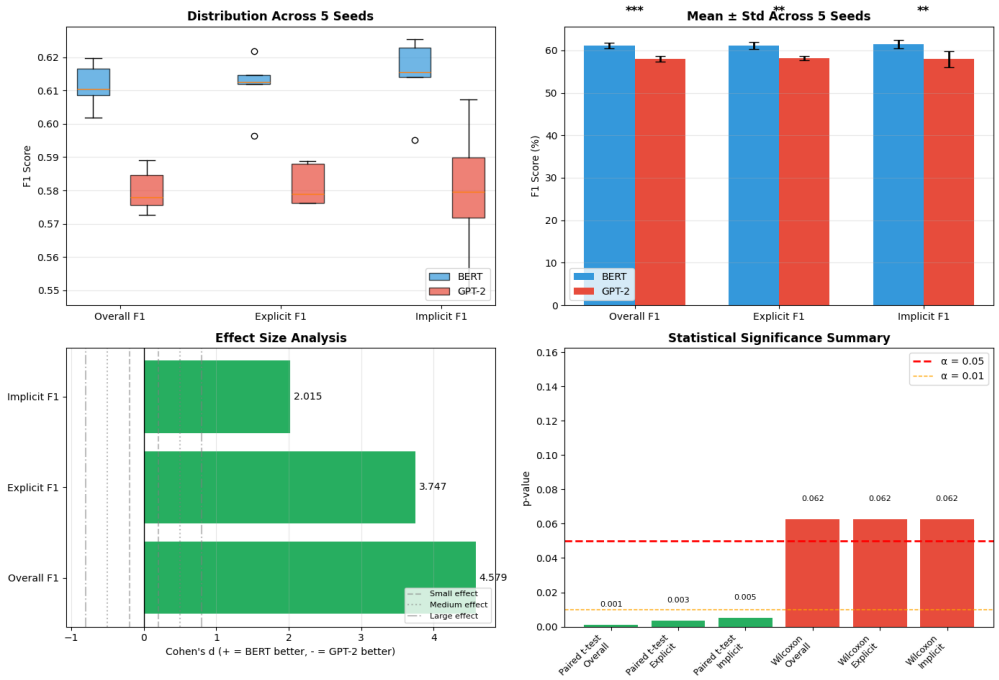


Figure 2. Multi-seed robustness summary (5 seeds, 3-epoch protocol): F1 distribution by seed (top left), mean±SD (top right), Cohen's d effect size (bottom left), and paired t-test / Wilcoxon p-values against $\alpha = 0.05$ and 0.01 (bottom right). Visualizes the statistics reported in Table 8.

Because this robustness check used a shorter training schedule and a different metric (F1) than the headline single-run accuracy figures in Table 4, the much smaller BERT−GPT-2 gap observed here (2–3 percentage points) should not be interpreted as contradicting the larger gap reported in Table 4 (32.3 points overall, 51.6 points implicit); rather, it indicates that the magnitude of the gap is sensitive to training budget, and that under-

trained GPT-2 models close much of the gap with BERT. This dependency is itself a relevant finding and is discussed further in Section V.

V. DISCUSSION

1. Bidirectional vs. Causal Attention

The 51.6-point implicit accuracy gap is consistent with bidirectional attention providing a substantial advantage in classification tasks that require holistic sentence understanding. However, BERT and GPT-2 in this study also differ in pretraining objectives, tokenizers, pooling strategies, and parameter counts, so the comparison cannot isolate attention direction as the sole explanation. In this study's data, the gap is nonetheless large enough (51.6 points) for architecture to be unlikely to be a negligible factor. Causal attention, optimized for next-token prediction, cannot integrate the dispersed implicit cues required by ABSA classification. GPT-2's 85.6% implicit error rate is unlikely to be explained by tuning alone and is consistent with a mismatch between the model's inductive bias and the task structure. However, this study cannot rule out contributions from the other architectural differences noted above. Based on this single autoregressive model, practitioners building production ABSA systems may wish to favor bidirectional encoders (BERT, RoBERTa, DeBERTa) over small-scale autoregressive models for classification subtasks [27], pending validation across additional autoregressive architectures.

2. Semantic Boundaries in Restaurant Discourse

Service-Food confusion reveals something important about the domain, not just the models. Aspect categories in restaurant reviews are not semantically discrete. Service quality routinely surfaces in food-related language, including delayed dish readiness, incorrect orders, and rushed food items. Model tuning alone cannot resolve this without a richer training signal. More principled remedies include targeted augmentation that generates sharp service-versus-food contrast examples, multi-label formulations that tolerate ambiguity, or graph-based representations that encode inter-category relationships [28].

3. Actionable Recommendations

These findings point to five concrete improvements that can be pursued independently or in combination. The most consequential is architectural: given GPT-2's 85.6% implicit error rate, this study suggests caution in using small-scale autoregressive models such as GPT-2 for ABSA classification without further architectural or training modification, and practitioners may consider bidirectional encoders such as BERT, RoBERTa, or DeBERTa as an alternative [27]; this recommendation is based on a single autoregressive architecture and would benefit from validation across additional models. Beyond architecture, the dominant failure pattern of Service-Food confusion calls for targeted data augmentation to generate training examples with sharply contrasting aspect boundaries, whether through paraphrasing or counterfactual generation. Class imbalance requires its own remedy: class-weighted loss or oversampling strategies, such as DeepSMOTE [29], are needed to provide minority classes with an adequate training signal, since both architectures currently fail nearly equally on rare classes. For deployment, BERT's 29 high-confidence errors are addressable via temperature scaling, whereas GPT-2's miscalibration is too severe for post hoc correction, and architectural replacement is the more appropriate response. Finally, implicit expressions that require domain-specific background knowledge unavailable from surface-form training benefit from external information injection via semantic wave enhancement [12] or knowledge graph augmentation [8], both of which have consistently shown gains on the restaurant benchmark.

VI. CONCLUSION

A total of 815 classification failures generated by BERT and GPT-2 on the SemEval 2016 Restaurant benchmark were analyzed and organized into a validated seven-category error taxonomy ($\kappa = 0.82$). Three findings stand out.

Service-Food confusion is the most common failure type (23.9%), a consequence of genuine semantic overlap in restaurant discourse rather than model weakness alone. Bidirectional attention is associated with substantially better performance in this comparison: BERT outperforms GPT-2 by 51.6 percentage points on implicit instances, and GPT-2 collapses to near-zero accuracy on eleven of twelve aspect categories. However, other architectural differences between the two models were not controlled for. Calibration analysis reinforces this asymmetry: BERT's confidence gap ($\Delta = 0.216$) shows its confidence scores are informative about correctness, while GPT-2's near-zero confidence gap ($\Delta = -0.001$) leaves its probability estimates unreliable for downstream use; calibration alone, however, does not establish deployment readiness. These results provide a concrete empirical foundation for targeted improvements: augmentation for confusable categories, class rebalancing for rare aspects, post-hoc calibration for BERT, and architectural replacement for GPT-2 in implicit detection settings. Future work should test knowledge-enhanced architectures and few-shot strategies for rare categories. The taxonomy and diagnostic framework are domain-agnostic and can be applied to any transformer-based ABSA system.

ACKNOWLEDGEMENT

The authors acknowledge support from Universiti Malaysia Terengganu (UMT) and Universitas Komputer Indonesia (Unikom) for resources and academic collaboration that enabled this research.

REFERENCES

[1]. A. Nazir, Y. Rao, L. Wu and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," IEEE Trans. Affect. Comput., vol. 15, no. 1, pp. 24–46, 2024.
[2]. B. Xing, W. Liao, H. Xie, and X. Zhao, "Aspect-based sentiment analysis: A comprehensive survey," J. King Saud Univ. Comput. Inf. Sci., vol. 36, no. 1, p. 101873, 2024.
[3]. H. Phan, T. Nguyen, D. Le, and M. Tran, "Unifying ABSA with BERT and multi-layered graph convolutional networks," Sci. Rep., vol. 14, p. 14589, 2024.
[4]. M. Pontiki et al., "SemEval-2016 Task 5: Aspect-based sentiment analysis," in Proc. SemEval, 2016, pp. 19–30.
[5]. G. S. Chauhan, R. Nahta, Y. K. Meena, and D. Gopalani, "Aspect-based sentiment analysis using deep learning approaches: A survey," Comput. Sci. Rev., vol. 49, p. 100576, 2023.
[6]. T. Jiang, J. Wang, and Y. Rao, "Attention-guided multi-task learning for aspect-based sentiment analysis," Neurocomputing, vol. 543, p. 126287, 2023.
[7]. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[8]. H. Wang, Y. Liu, and W. Zhu, "Knowledge-enhanced graph neural networks for implicit sentiment analysis," IEEE Trans. Knowl. Data Eng., vol. 36, no. 5, pp. 2134–2147, 2024.
[9]. S. Wu, Y. Xu, F. Wu, Z. Yuan, and Y. Huang, "Aspect-based sentiment analysis with type-aware graph convolutional networks," in Proc. ACL, 2023, pp. 3456–3467.
[10]. J. Zhou, H. Chen, X. Liu, and S. Wang, "MOIT: A novel task for mining opinions towards implicit targets," Eng. Appl. Artif. Intell., vol. 126, p. 106841, 2023.

[11]. X. Li, M. Chen, Y. Gao, and W. Zhang, "Implicit sentiment analysis with semantic-aware contrastive learning," Inf. Process. Manag., vol. 60, no. 4, p. 103365, 2023.
[12]. M. Zhang, F. Wu, W. Chen, X. Huang, and Y. Zhang, "Aspect-level implicit sentiment analysis model based on semantic wave and knowledge enhancement," J. Supercomput., vol. 80, pp. 22726–22747, 2024.
[13]. J. Fu et al., "An implicit aspect-based sentiment analysis method using supervised contrastive learning and knowledge embedding," Knowl.-Based Syst., vol. 302, p. 112386, 2024.
[14]. Z. Li, Y. Zou, C. Zhang, Q. Zhang, and Z. Wei, "Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training," in Proc. EMNLP, 2021, pp. 246–256.
[15]. H. Chen, J. Liu, X. Zhang, and S. Wang, "Graph-enhanced implicit aspect-level sentiment analysis based on multi-prompt fusion," Sci. Rep., vol. 15, p. 17321, 2025.
[16]. K. Scaria et al., "InstructABSA: Instruction learning for aspect-based sentiment analysis," in Proc. NAACL, 2024, pp. 1416–1430.
[17]. H. Luo, L. Ji, T. Li, and D. Jiang, "ChatGPT as a factual inconsistency evaluator for text summarization," in Findings NAACL, 2024, pp. 2456–2478.
[18]. T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld, "Errudite: Scalable, reproducible, and testable error analysis," in Proc. ACL, 2019, pp. 747–763.
[19]. M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in Proc. ACL, 2020, pp. 4902–4912.
[20]. H. Fei, F. Li, C. Li, S. Wu, J. Li, and D. Ji, "On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training," ACM Trans. Inf. Syst., vol. 41, no. 2, pp. 1–32, 2022.
[21]. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, 2017, pp. 1321–1330.
[22]. L. Clarté, B. Loureiro, F. Krzakala and L. Zdeborová, "Expectation consistency for calibration of neural networks," in Proc. UAI, 2023, pp. 443–453.
[23]. L. Frenkel and J. Goldberger, "Network calibration by temperature scaling based on predicted confidence," in Proc. EUSIPCO, 2022, pp. 1586–1590.
[24]. W. Chen et al., "A survey on imbalanced learning," Artif. Intell. Rev., vol. 57, pp. 1–65, 2024.
[25]. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Tech. Rep., 2019.
[26]. J. Cohen, "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.
[27]. L. Huang, C. Sun, and X. Qiu, "Towards unified aspect-based sentiment analysis with large language models," in Proc. ACL, 2024, pp. 2345–2358.
[28]. L. Shi, D. Han, J. Han, B. Qiao, and G. Wu, "Dependency graph enhanced interactive attention network for aspect sentiment triplet extraction," Neurocomputing, vol. 507, pp. 315–324, 2022.
[29]. D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," IEEE Trans. Neural Netw. Learn—Syst., vol. 34, no. 9, pp. 6390–6404, 2023.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

A Context-Aware Computer Vision-Based Model for ACL Injury Prevention in Female Youth Football: A PRISMA-based Review

Zodzwainashe Kudyachete
Nelson Mandela University
Gqeberha, EC, SA
s219189250@mandela.ac.za

Chinonye Leuna Obioha
Nelson Mandela University
Gqeberha, EC, SA
chinonye.obioha@mandela.ac.za

Darelle Van Greunen (Prof)
Nelson Mandela University
Gqeberha, EC, SA
darelle.vangreunen@mandela.ac.za

Abstract

Markerless computer vision is a low-cost, non-invasive solution for real-time monitoring of ACL injury risk in female youth football. Still, evidence on the design of such a system for resource-constrained developing countries remains fragmented across the human movement and digital technology literature. This paper synthesizes that evidence through a PRISMA-based systematic review of 30 studies published between 2019 and 2026. An integrated theoretical framework combining the Haddon Matrix and Dynamic Systems Theory was used to structure and translate the findings into design criteria. Findings which led to these design criteria were as follows: (1) markerless pose estimation can extract the kinematic parameters associated with ACL injury risk in female athletes; (2) computer vision can detect the high-risk non-contact football actions (landing, cutting, pivoting, and deceleration) in which non-contact ACL injuries commonly occur; (3) real-time decision support for coaches is both technically viable and necessary in resource-constrained environments; (4) existing systems are not calibrated for female youth populations, with mixed-gender and adult thresholds dominating the literature; and (5) practical implementation in controlled laboratory settings does not yet extend to outdoor, low-resource deployment. From these findings, six design criteria for a context-aware computer vision ACL prevention model for this population are derived, providing the evidence-based foundation for model design and expert validation in subsequent research.

Keywords: Computer Vision, Pose Estimation, ACL Injury Prevention, Female Youth Football, Real-Time Movement Analysis, Developing Countries

I. INTRODUCTION

Real-time, non-invasive computer vision presents a significant opportunity to address ACL injury risk in female youth football, yet translating this capability to resource-constrained developing-country contexts requires design evidence that the literature has not yet assembled in integrated form. Anterior cruciate ligament (ACL) injuries present challenges for female athletes, coaches, and football development programs, with women's football burdened by ACL injuries at four to six times higher than the rate at which male athletes would incur

injuries performing the same activities [1], [2]. These injuries have significant consequences on athletes beyond loss of playing time, including psychological trauma, football career disruption, and heightened long-term re-injury risk [3]. The challenge is more prevalent in developing countries, where resource constraints lead to limited access to medical facilities, substandard infrastructure, and systematic underinvestment in women's football development [4], [5]. Conventional injury prevention approaches, such as the FIFA 11+ and other conditioning programs, face limited adoption in resource-constrained settings due to a lack of specialist staff, attitudes towards injury prevention, and management prioritizing results over sustainable coaching practices [6]. Furthermore, conventional approaches tend to be limited by periodic assessments of injury risk that fail to capture more dynamic, habitual movement patterns that contribute to heightened ACL injury risk [7]. ICT and computer technologies, such as computer vision, machine learning, and artificial intelligence (AI), can help address these limitations through real-time monitoring and analysis on widely available mobile devices, allowing coaches and athletes to detect risk patterns earlier [8], [9]. The application of computer vision technologies in sports medicine to female athletes in developing countries remains an under-researched area [10], [11]. While existing studies focused more on controlled laboratory settings, there is a lack of studies that consider the real-time applications of computer vision technologies in practical sports environments [12]. Furthermore, research on the maturation- and gender-specific biomechanical risk factors affecting young female football athletes in relation to technological injury-prevention approaches remains largely limited [3], [13]. Therefore, examining whether evidence from existing computer vision-based systems can be transferred to sub-Saharan and other developing countries is an important gap to investigate. This paper addresses the problem that female youth football players in resource-constrained settings face a higher risk of suffering an ACL injury in high-risk movements; methods of monitoring that can detect the risk of such injuries are expensive and have been largely researched and validated under laboratory settings and on adult or mixed-gender populations. The evidence needed to design a context-appropriate alternative exists, but is dispersed across the biomechanics, sport technology, and health informatics literature. The aim of this paper is therefore to synthesize the evidence into design criteria for a context-aware, real-time computer vision model to monitor ACL injury risk in female youth football players in developing countries. The paper makes two contributions: first, it provides an integrated synthesis that connects computer vision feasibility, female biomechanics, and developing-country application evidence, and second, six evidence-based design criteria are derived from the synthesis that specify what a system must do, giving future model design and expert validation an evidential basis rather than an assumed one.

II. LITERATURE REVIEW

1. ACL Injury Burden in Female Youth Football

López-Valenciano et al. [1] and Pulici et al. [2] both highlight the significant challenges in women's football, with female players sustaining injuries at four to six times higher rates than their male counterparts during similar movement actions. There is thus a higher impact of consequences such as time loss, career disruption, psychological trauma, and a greater risk of re-injury [3]. The economic burden is evidently significant, with studies reporting ACL injury-related costs of up to €50 000, which is unsustainable for resource-constrained women's football development programs in developing countries where systematic under-investment in women's football remains [2], [4], [5]. This challenge is prevalent in youth football, where Sanchez-Sanchez et al. [3] found that female football players aged 13–20 face higher risk rates of 2.5 to 4.8 per 1000 hours of game time, with knee injuries contributing 15 to 25 percent of total injuries.

2. Biomechanical and Maturation-Specific Risk Factors

The link between ACL injury risk and movement quality during non-contact actions is well documented in [1], [13], [14], [15], and [16]. Sanchez-Sanchez et al. [3] found that non-contact actions such as cutting, pivoting, and landing were strongly linked to greater ACL injury risk. Hewett et al. [13] identified knee valgus angle, hip flexion, and neuromuscular control during dynamic tasks as predictors of ACL injury risk in female athletes. Anatomical differences, including a wider pelvis and increased quadriceps angle, as well as hormonal influences on ligament laxity across the menstrual cycle, further contribute to the heightened injury risk profile of female athletes [3]. A recent scoping review of prospective studies confirms that these kinematic measures remain the best-supported biomechanical predictors of primary ACL injury [17]. Injury risk is complicated by additional risk factors related to growth and maturation, with Sanchez-Sanchez et al. [3] and Wik [18] reporting evidence that the bodily changes that occur during adolescence lead to rapid musculoskeletal growth, which affects limb and muscle coordination during sports activities. Wik [18] further cites the need for more studies involving young female populations to address the significant gap in the literature.

3. Resource Constraints and ICT Adoption

The integration of ICT into sports injury prevention has progressed considerably, with technologies such as GPS, wearable biosensors, and motion capture showing promise in sports injury monitoring [8], [9]. The adoption of these technologies in women's football development programs remains limited in developing countries' contexts due to constraints in infrastructure, equipment, and technical expertise [4], [5], [11]. Bakare et al. [6] note that conventional injury prevention programs, such as FIFA 11+, face similarly limited uptake in resource-constrained settings due to the need for specialized expertise and effective program management. Mkumbuzi et al. [11] report the lack of medical staff in women’s football, particularly in developing countries, as a significant factor in higher injury risk in these settings.

4. Computer Vision in Sports Injury Prevention

Computer vision studies in sport show growing potential for pose estimation, movement classification, and feedback generation [8], [9], [12], [19], [20], [21]. Frameworks such as OpenPose, MediaPipe, and BlazePose enable non-invasive approaches to sports injury prevention, enabling real-time movement monitoring and biomechanical analysis [8], [9]. Recent applications have demonstrated the feasibility of pose estimation frameworks for assessing ACL injury risk indicators in athlete populations, with Blanchard et al. [12] and Ouf et al. [21] demonstrating computer vision-based knee valgus assessment in female athletes under controlled conditions. The application of machine learning alongside computer vision has further enabled the development of systems capable of classifying high-risk movement patterns in real time, supporting its potential for integration into training environments [20], [21]. However, existing implementations have largely been limited to controlled laboratory settings, with little consideration of the technical challenges of outdoor multi-player football environments or the constraints in developing countries [10], [12]. This gap establishes the need for the systematic review presented in this paper.

5. Integrated Theoretical Framework

The integrated framework combines the Dynamic Systems Theory (DST) and the Haddon Matrix, providing a lens for reporting the review findings. DST posits that injury risk is not static and arises from the dynamic relationship among a player, the movement action, and the environment [22], [23]. This underpins the identification of biomechanical and environmental parameters to be monitored in a computer vision system.

The Haddon Matrix provides a structured approach to mapping the review findings across the pre-event, event, and post-event phases of injury prevention [24]. The host, agent, and environment dimensions of the Haddon matrix are clearer categories for better understanding the evidence regarding biomechanical parameters, technical, and contextual requirements. This integrated framework will ensure that both technical and contextual considerations are taken into account in the review and the resulting model criteria for female youth football in resource-constrained settings.

III. METHODOLOGY

A systematic literature review (SLR) was conducted using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, thereby ensuring a transparent, standardized, and rigorous methodology [25]. The PRISMA framework was determined to be most appropriate for this study, as it is widely adopted for synthesizing evidence across the literature and is useful for identifying gaps worth noting for the development of new injury prevention interventions. The review protocol was not registered with PROSPERO or published in advance, as the review was conducted as part of a wider research study.

1. Search Strategy

The electronic databases utilized in the systematic search included PubMed, IEEE Xplore, and Elsevier ScienceDirect. The search terms used were constructed using Boolean operators and combined keywords as follows: (“computer vision” OR “motion analysis” OR “pose estimation”) AND (“ACL” OR “Anterior Cruciate Ligament” OR “knee injury” OR “lower body injuries”) AND (“female” OR “women” OR “gender” OR “youth”) AND (“football” OR “soccer” OR “sport”). Supplementary contextual terms were also included, such as “resource-constrained”, “developing countries”, “South Africa”, “South African”, and “sub-Saharan Africa”. Reference lists of included studies were also citation-chained to identify studies not returned by the database searches. The searches were restricted to peer-reviewed articles written in English. The initial search was conducted in September 2025 and was updated in July 2026 to capture work published after the original search. The search window was 2019 to 2026, chosen to capture more recent pose estimation and biomechanical research, during which technological capabilities have changed substantially. Studies published before 2019 were retained only where they were foundational to the field, and no later peer-reviewed study reported an equivalent finding. This was done to avoid the converse problem of recency bias, which may overlook previous solutions and approaches and their perceived shortcomings, which would be useful in identifying the model design criteria. Table 1 summarises the search strategy.

Table 1. Summary of Search Strategy

Component	Strategy
Databases	PubMed, IEEE Xplore, Elsevier ScienceDirect
Core search string	("computer vision" OR "motion analysis" OR "pose estimation") AND ("ACL" OR "anterior cruciate ligament" OR "knee injury") AND ("female" OR "women" OR "youth" OR "gender") AND ("football" OR "soccer" OR "sport")
Contextual terms	resource-constrained, developing countries, South Africa, sub-Saharan Africa
Date range	2019–2026. Initial search September 2025, updated July 2026. Pre-2019 studies were retained only where foundational and without a later peer-reviewed equivalent
Language and type	English-language peer-reviewed journal articles and peer-reviewed conference papers; official reports were retained only as contextual background and not counted in the 30-study SLR corpus

2. Inclusion and Exclusion Criteria

The inclusion criteria of the studies were that they: (1) reported on computer vision or AI-based approaches to injury risk monitoring and biomechanical analysis in sport; (2) mentioned ACL injury risk factors, injury prevention, or biomechanical parameters; (3) performed the study on datasets or human participants; and (4) were peer-reviewed journal articles or peer-reviewed conference papers. Reports from official governing bodies such as FIFA and other grey literature were consulted for contextual background, but were not included in the review corpus. The exclusion criteria applied to studies that: (1) exclusively focused on upper limb injuries; (2) focused on non-injury sport analytics; (3) used computer vision or AI-based approaches that focus on non-injury-based aspects of sport; and (4) reported the same finding for the same population using the same method as a study already included, in which case the more recent or larger study was retained.

3. Data Extraction and Quality Assessment

Data extraction was conducted using a standard template to capture the study objectives, data collection methodology, injury prevention methodology, biomechanical parameters analyzed, study findings, and study context. The quality of the studies was assessed using the Mixed Methods Appraisal Tool (MMAT) to account for the mix of qualitative and quantitative studies included. Appraisal was used to assess the confidence that could be placed in each study's contribution to the synthesis, thereby informing how much weight a study carried in theme construction rather than serving as an exclusion threshold; no study was excluded on quality grounds. Actual study appraisal results (i.e., scored appraisal of 30 studies according to all five criteria) were not included in this paper due to limitations disclosed in Section IV. Thematic coding was conducted manually. Stable identifiers (P1-P30, listed in Table 2) were created; all studies were independently coded against the fields collected in the extraction template. Firstly, codes were inductively extracted from the findings by manually reviewing each study. Then, possible themes were identified by grouping the codes in a second pass, with each study's computer vision capability as the main focus. Only themes for which evidence could be found in more than one study and that presented unique implications for computer vision system design were retained. This resulted in five themes as described in Section IV. These five themes were then mapped onto the integrated theoretical framework, which subsequently generated requirements. Technology-assisted coding using ‘Atlas.ti’ was a viable alternative and is planned for use in the primary data collection phase of the wider study.

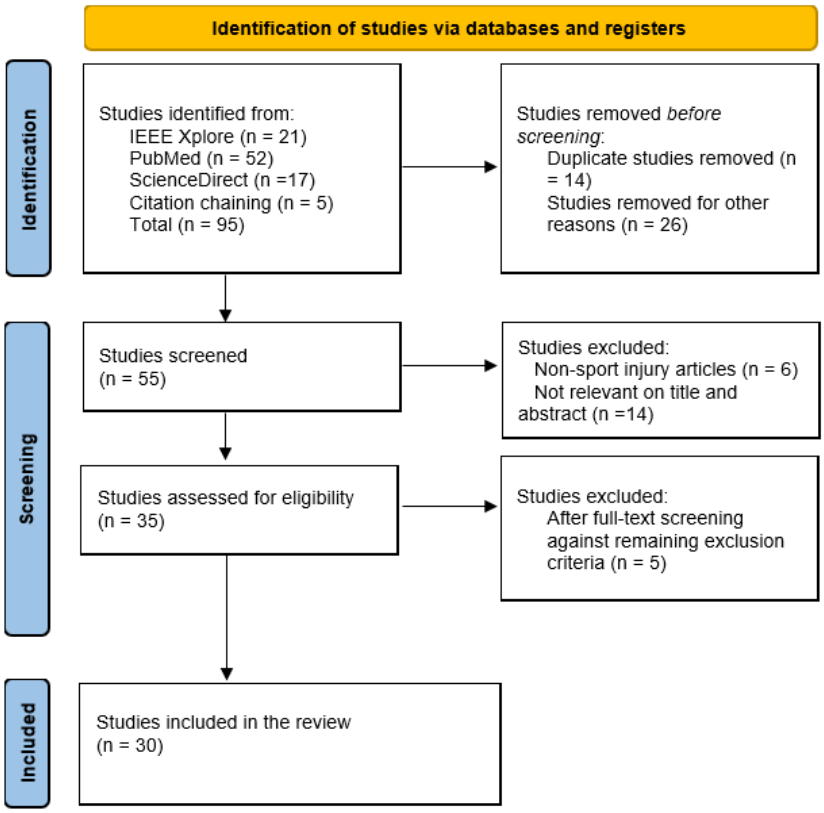


Figure 1. PRISMA Flow Diagram (95 identified; 55 screened; 25 excluded; 30 included)

Table 2. Summary of the articles selected for review

ID	Study	Primary focus
P1	López-Valenciano et al. (2021) [1]	Women's football injury epidemiology
P2	Pulici et al. (2023) [2]	Professional football injury burden and costs
P3	Sanchez-Sanchez et al. (2025) [3]	Young female soccer injuries
P4	Horan et al. (2023) [26]	Women's football injury incidence
P5	Zech et al. (2022) [27]	Sex differences in team-sport injury rates
P6	Parsons et al. (2021) [28]	Gendered environmental approach to ACL injury
P7	Bakare et al. (2021) [5]	Women's football injury prevention in South Africa
P8	Mkumbuzi et al. (2023) [11]	Sports medicine task sharing in grassroots women's football
P9	Omotayo et al. (2024) [10]	Computer vision research in Africa
P10	Hewett et al. (2005) [13]	Biomechanical predictors of ACL injury
P11	Lucarno et al. (2021) [29]	Video analysis of ACL injury mechanisms in female soccer
P12	Aiello et al. (2023) [30]	Football injury-inciting activities
P13	Wik (2022) [18]	Growth, maturation, and youth football injuries
P14	Johnson et al. (2022) [31]	Growth, exposure, and maturation-related injury risk

P15	Hewett and Bates (2017) [16]	Preventive biomechanics
P16	Hecksteden et al. (2023)[7]	Football injury forecasting with machine learning
P17	Naik et al. (2022) [8]	Computer vision in sport
P18	Zheng et al. (2025) [9]	Computer vision for football videos
P19	Blanchard et al. (2019) [12]	Computer vision for ACL risk in female athletes
P20	Ceballos-Laita et al. (2023) [19]	Computer vision knee valgus assessment
P21	Ouf et al. (2024) [21]	Computer vision injury prevention feedback
P22	Salian et al. (2024) [20]	Real-time joint-angle injury risk assessment
P23	Uhlrich et al. (2023) [32]	OpenCap smartphone-based movement dynamics
P24	Tan et al. (2023) [33]	Force-plate injury profiling methods
P25	Musat et al. (2024) [34]	AI-based sports injury risk prediction
P26	Molavi et al. (2026) [35]	AI in ACL injury prediction and prevention
P27	Chougule et al. (2026) [36]	Markerless versus marker-based lower-limb motion capture validity
P28	Huth et al. (2026) [37]	Machine learning injury prediction in elite women's football
P29	Chibhabha et al. (2026) [38]	Research agenda for COSAFA women's football
P30	Jimenez and Verhagen (2025) [39]	Athlete monitoring for injury prevention

IV. RESULTS

The searches returned 95 records: 52 from PubMed, 21 from IEEE Xplore, 17 from Elsevier ScienceDirect, and 5 identified through citation chaining. Forty records were removed before screening, comprising 14 duplicates and 26 records falling outside the date window or the sports injury domain. The remaining 55 studies were screened by title and abstract, of which 20 were excluded: 6 as not sport-injury related and 14 as not relevant to the review question. This left 35 reports assessed for full-text eligibility. A further 5 studies were excluded based on the criteria in Section III, resulting in the 30 studies synthesized in this review. PRISMA 2020 does not specify a minimum corpus size; the number of included studies is determined by the eligibility criteria rather than a predefined target [25]. The studies included quantitative, mixed-methods, and systematic reviews, with methodologies such as markerless pose estimation, machine-learning classifiers, smartphone video analysis, wearable sensors, and augmented reality. Five thematic areas were identified as follows: (1) markerless pose estimation for female ACL-relevant kinematics; (2) computer vision detection of high-risk football actions; (3) real-time and coach-facing decision support; (4) calibration gaps for female youth populations; and (5) field and resource-constrained deployment barriers.

1. Thematic Synthesis: Five Computer Vision Findings

Focusing on computer vision as the core of the synthesis, the study identified five distinct findings across the 30 selected articles. Table 3 maps each theme to the supporting articles and the main result.

Table 3. Thematic Synthesis of the 30 selected articles

Thematic areas	**Selected articles**	**Main result**

Theme 1: Markerless pose estimation for female ACL-relevant kinematics	P10, P12, P19-P23, P27	Frameworks such as OpenPose, MediaPipe, BlazePose, and OpenCap can extract knee valgus, knee flexion at landing, and joint angles that are linked to female ACL risk.
Theme 2: CV detection of high-risk football actions	P11, P12, P17, P18, P19-P22	Landing, cutting, pivoting, and sudden deceleration are the primary non-contact injury actions and are observable in video; detection and classification pipelines for these actions in football video have been validated.
Theme 3: Real-time, coach-facing decision support	P16, P21, P22, P25, P26, P28, P30	Computer vision outputs can be converted into real-time form feedback and joint-angle risk scores; monitoring must be decision support, not just data collection, especially where specialists are absent.
Theme 4: Calibration gap for female youth populations	P1-P6, P10, P13, P14, P19-P22, P26	Existing CV ACL systems use adult or mixed-gender thresholds and rarely account for maturation; sex-specific and growth evidence show that these thresholds are not transferable to female youth populations.
Theme 5: Field and resource-constrained deployment barriers	P7-P9, P17, P18, P23, P24, P27, P29, P30	CV evidence is overwhelmingly lab-based; outdoor football introduces occlusion, lighting, and multi-player tracking challenges, and video-only systems cannot fully replace kinetic tools; African CV infrastructure further constrains the field.

2. Theme 1: Markerless Pose Estimation for Female ACL-Relevant Kinematics

Pose estimation frameworks, such as OpenPose, MediaPipe, BlazePose, and the OpenCap system, can operate without physical markers by mapping keypoints to biomechanical kinematic parameters associated with female ACL risk. P10 [13] highlights knee valgus loading and reduced neuromuscular control as prospectively associated with ACL injury risk in female athletes, while P19 [12] and P20 [19] demonstrate that computer vision can measure knee valgus angle from video analysis with acceptable validity and reliability. P22 [20] extends this to real-time joint-angle scoring across multiple lower-limb joints. P23 [32] shows that smartphone video analysis is comparable with laboratory-grade motion capture for lower-limb kinematics, though still sensitive to camera placement and calibration. P24 [33] reports that out-of-lab systems estimate kinetic parameters, such as joint moments, through physics-based modeling or machine learning rather than measuring them directly, as force plates do, so kinetic estimates carry modeling assumptions that kinematic measures do not. Consequently, knee valgus, knee flexion at landing, hip and trunk control, and inter-limb asymmetry are observable in the video. Still, none of the studies establishes female-specific reference values against which to analyze them.

3. Theme 2: Computer Vision Detection of High-Risk Football Actions

The non-contact actions that cause ACL injuries (landing from a jump, side-step cutting, pivoting, and sudden deceleration) are observable through video analysis. P11 [29] demonstrates video-based identification of these actions through analysis of actual ACL injury incidents in professional women's football, and P12 [30] extends this to a wider range of football injury-inciting actions. P17 [8] and P18 [9] review developments in computer vision pipelines for football video in action detection and analysis, including player tracking, action recognition, and pose estimation using broadcast and mobile-camera footage. P21 [21] shows that a computer vision system can provide real-time feedback on movement form during physical activity, and P22 [20] shows that joint-angle-based risk scoring can run in a live athlete assessment context. A direct implication of the observed

studies is that the proposed model should recognize an injury-inducing action and then extract relevant kinematic parameters within the movement window, rather than continuously monitoring all frames. This way, fewer computational resources are required, which is key for minimal-resource deployment.

4. Theme 3: Real-Time, Coach-Facing Decision Support

P21 [21] and P22 [20] demonstrate that computer vision can provide real-time feedback that can inform corrective movement decisions. P30 [39] argues that monitoring must transition from data collection to actionable decision support. P16 [7] shows that machine-learning injury forecasting models can generate practical risk signals for coaching teams when trained on continuous monitoring data rather than one-off screening events, and P25 [34] reviews AI diagnostic tools for sports injury with findings that support clear risk categorization over vague scores. In the South African female youth football context, the argument is supported by P7 [5] and P8 [11], which show that coaches and non-specialist staff are the more practical end users of any injury-prevention tool. The model, therefore, needs to factor in color-coded risk severity levels, plain-language alerts, and post-session summaries that coaches can interpret without advanced medical expertise, while maintaining the underlying kinematic data for further analysis by medical experts.

5. Theme 4: Calibration Gap for Female Youth Populations

P1 [1] through P6 [28] establish that women and girls carry a distinct ACL injury burden and that the underlying risk elements are not gender-neutral. P10 [13] reports that valgus loading and neuromuscular control patterns differ between male and female athletes enough to invalidate the direct use of male-specific thresholds. P6 [28] extends this to the broader environmental and structural factors that shape female ACL risk, arguing against purely biological explanations. For youth athletes, P3 [3], P13 [18], and P14 [31] add an age dimension: rapid musculoskeletal growth during adolescence disrupts inter-limb coordination, changes landing mechanics as observed through computer vision, and shifts the risk profile over a single season. Yet the computer vision studies closest to this population (i.e., P19 [12], P20 [19], P21 [21], and P22 [20]) either use adult reference values, mixed-gender populations, or controlled laboratory settings that lack the context of a developing female player in an outdoor training environment. This calibration gap is a key finding of the review because it specifies the design requirement that the model should account for female- and age-appropriate thresholds.

6. Theme 5: Field and Resource-Constrained Deployment Barriers

P17 [8] and P18 [9] identify occlusion, variable lighting, multi-player tracking, and camera quality as the main barriers to effective outdoor/live training computer vision. P24 [33] reviews out-of-lab portable sensing for ACL assessment and finds that most approaches identified remain proofs of concept, lacking clinical validation or widespread adoption. In contrast, laboratory assessment remains expensive and inaccessible to most players. P9 [10] shows that African computer vision research is constrained by limited local datasets, infrastructure gaps, and barriers to collaboration that affect both model training and practical deployment. These constraints are especially evident in the South African women's football environment, as validated by P7 [5] and P8 [11]. Despite these barriers, P23 [32] provides relevant evidence through the OpenCap system, demonstrating that multi-smartphone video analysis can achieve valid lower-limb kinematics comparable to laboratory motion capture, with a guided setup rather than specialist calibration. P30 [39] argues that injury monitoring approaches must connect field data to practical injury-prevention decisions, which the model should factor in for coaches in grassroots women's football.

7. Design Criteria and Integrated Framework Mapping

The five thematic findings, analyzed using the Haddon Matrix and the DST framework, yield six evidence-based design criteria for a context-aware computer vision model aimed at preventing ACL injuries. Each design criterion relates to the thematic evidence and outlines a requirement that current systems fail to meet for female youth in resource-constrained contexts. Together, they define what such a system must do to be technically valid and practically deployable in resource-constrained women's football programs. The criteria strictly describe what the evidence indicates such a system would need to do, where future research will assess the effectiveness of these criteria. Read through the integrated theoretical framework; these five findings translate into the design requirements summarised in Table 4.

Table 4. Evidence-based design criteria for the proposed model

Criterion	**Practical meaning**	**Supporting selected articles**	**Gap in existing work**
Female- and youth-specific baseline	Interpret risk using player-specific, female-specific, and age-aware information.	P1-P6, P10, P13-P15, P26	Existing CV systems use adult or mixed-gender reference thresholds; no implementation reviewed builds female- or youth-specific baselines for this population.
High-risk action detection	Identify landing, cutting, pivoting, and sudden deceleration during football activities.	P11-P12, P19-P22	Action detection pipelines exist for general sport but have not been tested in outdoor female youth football environments in developing countries.
Kinematic parameter set	Extract knee valgus, knee flexion, hip/trunk control, and asymmetry from the video.	P10-P12, P15, P20, P23-P24, P27	Parameters are technically extractable, but no reviewed system validates the full set against female-specific norms outside a laboratory.
Low-cost technical setup	Use smartphone or consumer-camera input, simple calibration, and limited-compute processing.	P9, P17-P18, P21-P23, P27	Smartphone-based systems exist but have not been deployed or validated in outdoor multi-player football settings in Africa or comparable LMIC contexts.
Actionable decision support	Return clear severity levels and practical feedback, not just raw data.	P16, P25, P26, P28, P30	Existing outputs target analysts or clinicians; no reviewed system produces coach-facing, plain-language risk feedback for non-specialist users in low-resource settings.
Continuous monitoring	Store session trends and update risk profiles as the player and season change.	P13-P16, P24, P28, P30	All reviewed CV systems operate as one-off assessments; none continuously track risk over a season in female youth football.

8. Limitations

None of the 30 studies directly validates a computer vision ACL injury-prevention model for female youth football players in resource-constrained settings, so the design criteria are drawn from adjacent evidence and still need to be tested in the target context. The corpus also combines primary studies with systematic and narrative reviews, so some primary evidence is counted twice: once on its own and once inside a review that includes it. The 30 studies are therefore best read as the evidence available at this intersection, rather than as 30 independent samples. The studies vary in design, participant groups, and measurement tools, which is why the evidence was synthesized thematically and not through a meta-analysis.

Three further limitations relate to the search and the appraisal. Only PubMed, IEEE Xplore, and ScienceDirect were searched, so relevant work indexed in Scopus, Web of Science, or the ACM Digital Library may have been missed. One reviewer also performed the searching, screening, extraction, and appraisal, which raises the risk of selection and extraction bias; co-author review of the protocol, coding frame, and final corpus reduces this risk but does not remove it. MMAT results and the full database-specific filters are not reported here, as neither fits within the prescribed length of this conference paper; both will be reported in the extended version of this study.

Lastly, the corpus is weighted towards work published between 2021 and 2024, with four of the 30 studies published in 2026. Two studies published before 2019 were retained under the rule set out in Section III.1. The prospective evidence linking knee abduction loading to ACL injury in female athletes has not been replicated in a comparable cohort since 2005, reflecting the state of the evidence base.

V. CONCLUSION

This PRISMA-based systematic review synthesized 30 selected studies to establish the evidence base for a context-aware computer vision model for ACL injury prevention in female youth football. The paper makes two contributions. First, a five-theme synthesis that connects computer vision feasibility, biomechanics, and evidence on developing-country deployment in resource-constrained contexts. Second, six evidence-based design criteria that specify what a female-youth-specific, resource-appropriate computer vision ACL prevention system must do across player profiling, action detection, kinematic extraction, risk interpretation, and deployment. These criteria are designed to be taken forward in future research on model design and expert validation work with computer vision specialists, human movement experts, and football coaches. This constitutes the next phase of the broader research of which this paper forms a part. The paper contributes to the broader context of applied AI for healthcare and to the Sustainable Development Goals for women in developing countries.

ACKNOWLEDGEMENT

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article. AI Declaration. The authors acknowledge the use of generative AI tools to support language editing and structural refinement. The authors reviewed, edited, and approved the final content of this paper. The primary study has been submitted to the Ethics Committee at Nelson Mandela University for ethics approval.

REFERENCES

[1] A. López-Valenciano et al., “Injury Profile in Women’s Football: A Systematic Review and Meta-Analysis,” Sports Medicine, vol. 51, no. 3, pp. 423–442, Mar. 2021, doi: 10.1007/s40279-020-01401-w.

[2] L. Pulici, D. Certa, M. Zago, P. Volpi, and F. Esposito, "Injury Burden in Professional European Football (Soccer): Systematic Review, Meta-Analysis, and Economic Considerations," Clinical Journal of Sport Medicine, vol. 33, no. 4, pp. 450–457, Jul. 2023, doi: 10.1097/JSM.0000000000001107.
[3] J. Sanchez-Sanchez, J. Raya-González, V. Martín, and A. Rodríguez Fernández, "Understanding Injuries in Young Female Soccer Players: A Narrative Review on Incidence, Mechanism, Location Risk Factors, and Preventive Strategies," Applied Sciences, vol. 15, no. 3, p. 1612, Feb. 2025, doi: 10.3390/app15031612.
[4] Commission for Gender Equality, "Gender Inequality in Sport: Insights into the Women's National Soccer Team – Banyana Banyana 2024," Commission for Gender Equality, 2024. Accessed: Aug. 27, 2025. [Online]. Available: https://www.parliament.gov.za/storage/app/media/Docs/tpap/1ae48a40-a643-4c30-8270-9cc7ef757b94.pdf
[5] U. Bakare, B. Olivier, C. Brandt, and L. Goldwana, "Injury prevention knowledge, beliefs, and practices among women's football teams in South Africa," South African Journal of Sports Medicine, vol. 33, no. 1, pp. 1–6, Jul. 2021, doi: 10.17159/2078-516X/2021/v33i1a9505.
[6] U. Bakare, "The development of an implementation guide for the FIFA 11+ injury prevention program among female football players in South Africa," 2022. Accessed: Aug. 30, 2025. [Online]. Available: https://hdl.handle.net/10539/35652
[7] A. Hecksteden, G. P. Schmartz, Y. Egyptien, K. Aus der Fünten, A. Keller, and T. Meyer, "Forecasting football injuries by combining screening, monitoring and machine learning." Science & medicine in football, vol. 7, no. 3, pp. 214–228, Aug. 2023, doi: 10.1080/24733938.2022.2095006.
[8] B. T. Naik, M. F. Hashmi, and N. D. Bokde, "A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions," Applied Sciences, vol. 12, no. 9, p. 4429, Apr. 2022, doi: 10.3390/app12094429.
[9] F. Zheng, D. Z. Al-Hamid, P. H. J. Chong, C. Yang, and X. J. Li, "A Review of Computer Vision Technology for Football Videos," Information, vol. 16, no. 5, p. 355, Apr. 2025, doi: 10.3390/info16050355.
[10] A.-H. Omotayo et al., "The State of Computer Vision Research in Africa," Journal of Artificial Intelligence Research, vol. 81, pp. 43–69, Sep. 2024, doi: 10.1613/jair. 1.16653.
[11] N. S. Mkumbuzi, A. Massey, S. K. Lubega, B. Sorowen, and E. M. Chisati, "FIFA football nurse – A task sharing approach in sports and exercise medicine practice in grassroots women's football in low- and middle- income settings. A study protocol for a cluster randomized controlled trial," PLoS One, vol. 18, no. 9, p. e0278428, Sep. 2023, doi: 10.1371/journal.pone.0278428.
[12] N. Blanchard, K. Skinner, A. Kemp, W. Scheirer, and P. Flynn, "'Keep Me In, Coach!': A Computer Vision Perspective on Assessing ACL Injury Risk in Female Athletes," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Jan. 2019, pp. 1366–1374. doi: 10.1109/WACV.2019.00150.
[13] T. E. Hewett et al., "Biomechanical Measures of Neuromuscular Control and Valgus Loading of the Knee Predict Anterior Cruciate Ligament Injury Risk in Female Athletes: A Prospective Study," Am. J. Sports Med., vol. 33, no. 4, pp. 492–501, Apr. 2005, doi: 10.1177/0363546504269591.
[14] D. A. Padua, S. W. Marshall, M. C. Boling, C. A. Thigpen, W. E. Garrett, and A. I. Beutler, "The Landing Error Scoring System (LESS) Is a Valid and Reliable Clinical Assessment Tool of Jump-Landing Biomechanics: The JUMP-ACL Study," Am. J. Sports Med., vol. 37, no. 10, pp. 1996–2002, Oct. 2009, doi: 10.1177/0363546509343200.
[15] T. A. Donelon, J. Edwards, M. Brown, P. A. Jones, J. O'Driscoll, and T. Dos'Santos, "Differences in Biomechanical Determinants of ACL Injury Risk in Change of Direction Tasks Between Males and Females: A Systematic Review and Meta-Analysis," Sports Med. Open, vol. 10, no. 1, p. 29, Apr. 2024, doi: 10.1186/s40798-024-00701-z.

[16] T. E. Hewett and N. A. Bates, "Preventive Biomechanics: A Paradigm Shift With a Translational Approach to Injury Prevention," Am. J. Sports Med., vol. 45, no. 11, pp. 2654–2664, Sep. 2017, doi: 10.1177/0363546516686080.
[17] R. K. Straub and C. M. Powers, "Biomechanical predictors of primary ACL injury: A scoping review of prospective studies," Gait Posture, vol. 116, pp. 22–29, Feb. 2025, doi: 10.1016/j.gaitpost.2024.11.013.
[18] E. H. Wik, "Growth, maturation and injuries in high-level youth football (soccer): A mini review," Front. Sports Act. Living, vol. 4, p. 975900, Nov. 2022, doi: 10.3389/fspor. 2022.975900.
[19] L. Ceballos-Laita, X. Marimon, A. Masip-Alvarez, S. Cabanillas-Barea, S. Jiménez-del-Barrio, and A. Carrasco-Uribarren, "A Beta Version of an Application Based on Computer Vision for the Assessment of Knee Valgus Angle: A Validity and Reliability Study," Healthcare, vol. 11, no. 9, p. 1258, Apr. 2023, doi: 10.3390/healthcare11091258.
[20] A. Salian, B. Ghosh, H. Shetye, P. Kapadia, and K. Srivastava, "Real-Time Injury Risk Assessment in Athletes Based on Relative Joint Angles," in 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/ESCI59607.2024.10497417.
[21] M. A. Ouf, R. M. Amin, M. M. Sabry, A. H. Ali, Y. A. Hamdan, and S. N. Saleh, "VISION-PF: A Computer Vision Approach to Injury Prevention in Physical Fitness," in 2024 International Telecommunications Conference (ITC-Egypt), IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/ITC-Egypt61547.2024.10620528.
[22] N. F. N. Bittencourt, W. H. Meeuwisse, L. D. Mendonça, A. Nettel-Aguirre, J. M. Ocarino, and S. T. Fonseca, "Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept," Br. J. Sports Med., vol. 50, no. 21, pp. 1309–1314, Nov. 2016, doi: 10.1136/BJSPORTS-2015-095850.
[23] P. S. Glazier and K. Davids, "Constraints on the Complete Optimization of Human Motion," Sports Medicine, vol. 39, no. 1, pp. 15–28, Sep. 2009, doi: 10.2165/00007256-200939010-00002.
[24] W. Haddon, "A Logical Framework for Categorizing Highway Safety Phenomena and Activity," The Journal of Trauma: Injury, Infection, and Critical Care, vol. 12, no. 3, pp. 193–207, Mar. 1972, doi: 10.1097/00005373-197203000-00002.
[25] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, Mar. 2021, doi: 10.1136/BMJ.N71.
[26] D. Horan, F. Büttner, C. Blake, M. Hägglund, S. Kelly, and E. Delahunt, "Injury incidence rates in women's football: a systematic review and meta-analysis of prospective injury surveillance studies," Br. J. Sports Med., vol. 57, no. 8, pp. 471–480, Apr. 2023, doi: 10.1136/bjsports-2021-105177.
[27] A. Zech et al., "Sex differences in injury rates in team-sport athletes: A systematic review and meta-regression analysis," J. Sport Health Sci., vol. 11, no. 1, pp. 104–114, Jan. 2022, doi: 10.1016/j.jshs.2021.04.003.
[28] J. L. Parsons, S. E. Coen, and S. Bekker, "Anterior cruciate ligament injury: towards a gendered environmental approach," Br. J. Sports Med., vol. 55, no. 17, pp. 984–990, Sep. 2021, doi: 10.1136/bjsports-2020-103173.
[29] S. Lucarno et al., "Systematic Video Analysis of Anterior Cruciate Ligament Injuries in Professional Female Soccer Players," Am. J. Sports Med., vol. 49, no. 7, pp. 1794–1802, Jun. 2021, doi: 10.1177/03635465211008169.
[30] F. Aiello, F. M. Impellizzeri, S. J. Brown, A. Serner, and A. McCall, "Injury-Inciting Activities in Male and Female Football Players: A Systematic Review," Sports Medicine, vol. 53, no. 1, pp. 151–176, Jan. 2023, doi: 10.1007/s40279-022-01753-5.

[31] D. M. Johnson, S. P. Cumming, B. Bradley, and S. Williams, "The influence of exposure, growth and maturation on injury risk in male academy football players," J. Sports Sci., vol. 40, no. 10, pp. 1127–1136, May 2022, doi: 10.1080/02640414.2022.2051380.
[32] S. D. Uhlrich et al., "OpenCap: Human movement dynamics from smartphone videos," PLoS Comput. Biol., vol. 19, no. 10, p. e1011462, Oct. 2023, doi: 10.1371/journal.pcbi.1011462.
[33] T. Tan et al., "A scoping review of portable sensing for out-of-lab anterior cruciate ligament injury prevention and rehabilitation," NPJ Digit. Med., vol. 6, no. 1, Mar. 2023, doi: 10.1038/S41746-023-00782-2.
[34] C. L. Musat et al., "Diagnostic Applications of AI in Sports: A Comprehensive Review of Injury Risk Prediction Methods," Diagnostics, vol. 14, no. 22, p. 2516, Nov. 2024, doi: 10.3390/diagnostics14222516.
[35] M. R. Molavi, M. Mottaghitalab, P. Samaei, and H. Zarei, "Artificial intelligence in ACL injury prediction and prevention: a systematic review," Journal of Orthopedic Surgery and Research 2026 21:1, vol. 21, no. 1, p. 330, Apr. 2026, doi: 10.1186/S13018-026-06825-0.
[36] A. Chougule et al., "Accuracy and Validity of 3D Markerless Motion Capture Compared to Marker-Based Systems for Lower-Limb Biomechanical Assessment: A Systematic Review," Sensors, vol. 26, no. 12, p. 3956, Jun. 2026, doi: 10.3390/s26123956.
[37] M. Huth et al., "Injury prediction in elite women's football: an integrative machine learning-based decision-support framework," NPJ Digit. Med., Jul. 2026, doi: 10.1038/s41746-026-02937-3.
[38] F. Chibhabha, F. B. B. Ndiaye, M. O. Ngcobo, and N. S. Mkumbuzi, "Setting the research agenda for Council of Southern Africa Football Associations (COSAFA) women's football: stakeholder perspectives," BMC Sports Sci. Med. Rehabil., vol. 18, no. 1, p. 254, Apr. 2026, doi: 10.1186/s13102-026-01692-y.
[39] C. Jimenez and E. Verhagen, "Reimagining athlete monitoring for true indicative injury prevention," BMJ Open Sport Exerc. Med., vol. 11, no. 2, p. e002479, Jun. 2025, doi: 10.1136/BMJSEM-2025-002479.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Synthetic-to-Real Camera Contamination Classification Using Modern Deep Neural Networks

János Hollósi
Department of Artificial Intelligence and High Performance Computing
Széchenyi István University
Győr, Hungary
hollosi.janos@sze.hu
ORCID:0000-0001-9002-9410

Abstract
Camera lens contamination from mud, water droplets, dust, or other environmental factors can significantly degrade the performance of vision-based perception systems in autonomous vehicles and mobile robots. This paper presents a comprehensive comparison of fifteen state-of-the-art image classification architectures for automatic camera contamination recognition formulated as a binary classification task. The evaluated models include convolutional neural networks (CNNs), vision transformers, and hybrid architectures. All models are trained exclusively on a synthetic dataset generated with the aiSim autonomous driving simulator. To evaluate synthetic-to-real domain transfer, the trained models are tested on a manually annotated subset of the real-world WoodScape dataset, which contains both clean and naturally mud-contaminated images. The experimental results demonstrate that synthetic training data can effectively support real-world camera contamination recognition despite the domain gap between simulation and reality. Swin-T achieved the highest F1-score, while MaxViT-T obtained the highest ROC-AUC and PR-AUC values. Furthermore, the results indicate that network architecture has a considerably greater influence on cross-domain generalization than model complexity alone. These findings demonstrate the feasibility of using synthetic datasets to develop reliable camera-contamination classifiers and provide practical guidance for selecting suitable image-classification architectures for autonomous driving applications.

Keywords: camera contamination, image classification, synthetic data, domain adaptation, autonomous driving

I. INTRODUCTION

Vision-based perception is a key component of modern advanced driver assistance systems (ADAS) and autonomous vehicles, supporting tasks such as object detection, lane detection, and traffic sign recognition. However, adverse environmental conditions, particularly camera lens contamination caused by mud, water droplets, dust, or insects, can significantly degrade image quality and reduce the reliability of downstream computer vision algorithms. Previous studies have demonstrated the negative impact of such contamination on both image quality and object detection performance, highlighting the importance of reliable camera condition assessment [1-3]. In our previous work, we investigated the impact of synthetic mud contamination on object detection using aiSim-generated datasets [4] and later extended this analysis to naturally occurring contamination [5]. Although these studies demonstrated the challenges posed by camera contamination, they focused on object detector robustness rather than identifying contaminated images. In this paper, we address this problem as a binary image classification task by comparing fifteen state-of-the-art CNN, transformer-based, and hybrid architectures. All models are trained exclusively on synthetic images generated with aiSim and evaluated on manually annotated real-world images from the WoodScape dataset, enabling a comprehensive assessment of synthetic-to-real generalization for automatic camera contamination detection. The remainder of this paper is organized as follows. Section II reviews related work. Section III introduces the aiSim and WoodScape datasets.

Section IV describes the investigated neural network architectures and the experimental methodology. Section V presents the experimental results, while Section VI concludes the paper and outlines future research directions.

II. RELATED WORK

Reliable perception is essential for autonomous vehicles and advanced driver assistance systems, yet the performance of vision-based systems can be severely degraded by contaminants adhering to the camera lens. Consequently, recent research has focused on three main directions: image restoration techniques, contamination detection methods, and synthetic data generation to improve model robustness. One research direction aims to restore degraded images by removing the visual effects of contaminants. He et al. [6] proposed an attentive convolutional neural network for removing adherent mist and raindrops from a single image, demonstrating that restoring image quality significantly improves subsequent vision tasks. Rather than identifying contamination, their approach reconstructs a clean image from a degraded observation, making it suitable as a preprocessing step for downstream perception algorithms. One of the most influential works in this area is SoilingNet [7], which formally introduced the camera soiling detection problem for automotive surround-view cameras. The authors proposed both image-level and tile-level contamination detection using convolutional neural networks. They introduced a large-scale dataset containing multiple types of contamination, including mud, dust, water, and snow. To further improve robustness, they also investigated GAN-based data augmentation by generating synthetic contaminations that complemented the training data [8]. Building upon this work, TiledSoilingNet [9] introduced a coverage-based formulation that estimates contamination severity at the tile level rather than producing only a global image label, enabling a more detailed assessment of contaminated regions. Recent studies have also focused on lightweight deep learning models suitable for embedded automotive platforms. SoildNet [10] is a computationally efficient convolutional neural network employing dynamic group convolutions and channel reordering to reduce computational complexity while maintaining competitive detection accuracy. Unlike earlier approaches that emphasize large CNN architectures, their work demonstrates that contamination detection can also be achieved with compact models designed for real-time execution on resource-constrained hardware. Kim et al. [11] presented a YOLOv5-based droplet-contamination detector trained on a manually annotated dataset comprising approximately 4,000 images. In addition to comparing multiple YOLOv5 variants, the authors integrated the detector with a transparent heater-based cleaning mechanism, demonstrating an end-to-end automatic contamination detection and cleaning system suitable for practical deployment. These studies primarily formulate contamination detection as an object detection problem, in which individual droplets are localized with bounding boxes rather than assessing the overall contamination state of the image. Another important line of research addresses raindrop detection. Kurihata et al. [12] introduced a dedicated windshield raindrop dataset together with a lightweight gradient-based detection algorithm that avoids computationally expensive deep neural networks while maintaining real-time performance. Their work illustrates that carefully designed, handcrafted methods can still provide practical solutions for specific types of contamination. However, they are generally less flexible than modern learning-based approaches across diverse environmental conditions. Recent studies have also investigated the use of synthetic data and its transferability to real-world perception tasks in autonomous driving. Silva et al. [13] systematically investigated how different aspects of synthetic image generation, including geometry, materials, lighting, rendering noise, and post-processing, affect the synthetic-to-real generalization of semantic segmentation models in autonomous driving. Miller et al. [14] proposed EdgeNet, an EfficientNetV2-based end-to-end autonomous driving architecture that uses edge detection and synthetic CARLA data for pretraining, demonstrating improved synthetic-to-real transfer and reduced training time on real-world driving data. Xu et al. [15] proposed a scenario-driven synthetic data generation framework that systematically integrates ODD and OEDR specifications to produce controllable autonomous driving datasets under adverse conditions, demonstrating that the resulting synthetic data can support the evaluation of perception robustness and improve model generalization.

III. DATASETS

The synthetic dataset used in this study was generated with the aiSim autonomous driving simulator developed by aiMotive. The simulator provides a realistic virtual environment for autonomous driving research, supporting reproducible traffic scenarios under controlled conditions, as well as camera contamination simulation. The dataset contains images captured in six urban and suburban environments under sunny, rainy, and snowy conditions. All RGB images have a resolution of 1280×720 pixels. The original dataset consists of 7,288 clean images. For each image, t10 additional mud-contaminated versions were generated using aiSim's built-in contamination model, resulting in different contamination patternsand spatial distributions. Since this work addresses binary image classification, one contaminated version was randomly selected for each clean image. The final dataset , htherefore, contains7,288 clean and 7,288 contaminated images. Representative examples are shown in Fig. 1. The dataset was divided into training, validation, and test subsets using a 70%, 15%, and 15% split. To prevent data leakage, clean and contaminated versions of the same source image were never included simultaneously, and the split was designed to minimize scene overlap between subsets. This strategy provides a more reliable evaluation of model generalization.



Figure 1. Samples from the synthetic aiSim dataset. The first row contains clean images, while the second row presents examples with simulated mud contamination.

To evaluate synthetic-to-real generalization, the publicly available WoodScape dataset [16] was used. WoodScape is a real-world automotive fisheye image dataset containing more than 10,000 images with annotations for several perception tasks. Among the publicly available datasets examined in this study, it was the only one containing a sufficiently large number of images with naturally occurring mud contamination, making it suitable for evaluation. As WoodScape does not provide camera contamination annotations, all images were manually inspected. A total of 3,698 images with visible mud contamination were identified, while an additional 3,698 clean images were randomly selected and visually verified. The resulting evaluation dataset, therefore, contains 7,396 images, evenly distributed between the two classes. Representative examples are shown in Fig. 2. Unlike the synthetic aiSim dataset, WoodScape was used exclusively for testing. No images were included in the training or validation process, allowing evaluation of how well models trained solely on synthetic data generalize to real-world camera contamination.

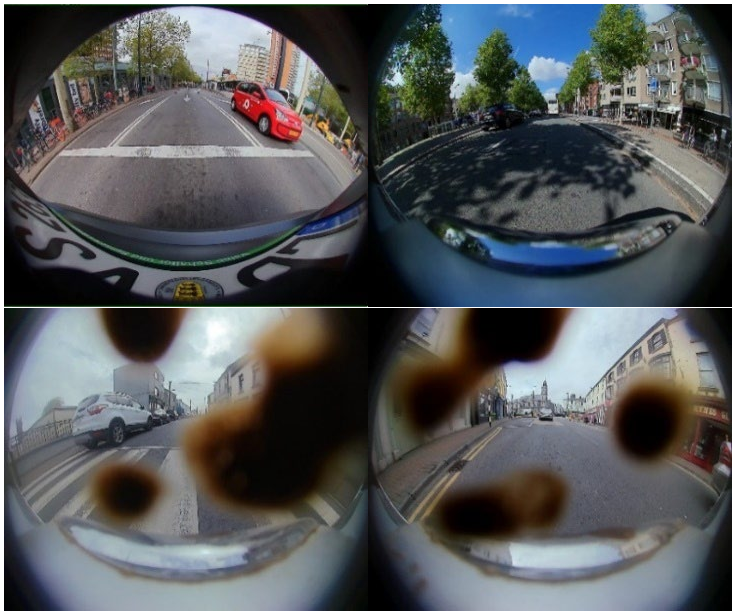


Figure 2. Samples from the WoodScape dataset. The first row contains clean images, while the second row shows images with naturally occurring mud contamination on the camera lens.

IV. METHODOLOGY

To investigate synthetic-to-real generalization, fifteen state-of-the-art image classification models representing CNN, Transformer, and hybrid architectures were evaluated. Table 1 summarizes the evaluated models: CAFormer-S18 [17], CoAtNet-0 [18], ConvNeXtV2-Tiny [19], DeiT-Small [20], ECA-NFNet-L0 [21], EfficientNetV2-S [22], LeViT-128 [23], MaxViT-Tiny [24], MobileNetV3-Large [25], PoolFormer-S12 [26], PVTv2-B2 [27], RegNetY-032 [28], ResNet50 [29], Swin-Tiny [30], and ViT-Base [31]. All models were initialized with ImageNet-pretrained weights and fine-tuned after replacing the original classification layer with a two-class output layer. To ensure a fair comparison, identical training settings were used for all models. Input images were resized to 224×224 pixels. Training employed the AdamW optimizer [32] with an initial llearning rate of 1×10 üssü eksi 4 son üst simge, weight of 1×10 üssü , eksi 4 son üst simge, weight decay of 1×10 üssü , eksi 4 son üst simge, weiweight decay 1×10^{-4}, weigdecay of 1×10^{-4}, cross-entropy loss, a batch size of 8, and 15 training epochs. The learning rate followed a cosine annealing schedule ($T_{\max} = 15$). Data augmentation consisted of horizontal flipping ($p = 0.5$) and color jittering of brightness, contrast, saturation, and hue. All experiments were implemented using PyTorch 1.12 [33]. Each model was trained on the synthetic aiSim dataset, while the validation set was used for model selection. Performance was then evaluated on the held-out synthetic test set and the independently constructed real-world WoodScape dataset, enabling direct comparison of synthetic-to-real generalization under identical training conditions.

Table 1. Overview of the evaluated image classification models

Category	Model	Parameters
CNN	MobileNetV3-Large [25]	4,204,594
	RegNetY-3.2GF [28]	17,926,364
	ECA-NFNet [21]	21,843,534
	EfficientNetV2 [22]	22,151,882

	ResNet-50 [29]	23,512,130
	ConvNeXt V2-T [19]	27,868,034
Hybrid	LeViT-128 [23]	8,445,476
	CAFormer-S18 [17]	24,296,754
	CoAtNet-0 [18]	26,668,100
	MaxViT-T [24]	30,404,554
Transformer	PoolFormer-S12 [26]	11,403,202
	DeiT-S [20]	21,666,434
	PVTv2-B2 [27]	24,850,882
	Swin-T [30]	27,520,892
	ViT-Base [31]	85,800,194

V. EXPERIMENTAL RESULTS

The proposed approach was evaluated by training all classification models exclusively on the synthetic aiSim dataset and subsequently evaluating them on the real-world WoodScape dataset. Consequently, the reported results quantify each architecture's ability to generalize across the synthetic-to-real domain gap, rather than its performance on data from the same distribution. Fig. 3 summarizes the overall performance of the evaluated architectures. Despite being trained solely on synthetic images, several models achieved excellent performance on real-world camera contamination, demonstrating that the synthetic contamination generated by aiSim provides sufficiently representative visual characteristics for learning transferable feature representations. Among all evaluated models, Swin-T achieved the highest classification performance with an F1-score of 0.965, followed by CoAtNet-0 (0.945), PoolFormer-S12 (0.923), and MaxViT-T/ECA-NFNet (0.911). In contrast, DeiT-S obtained the lowest performance, indicating that the ability to generalize to real contamination strongly depends on the selected network architecture. The relationship between model complexity and classification performance further confirms that increasing the number of trainable parameters alone does not guarantee improved generalization. Although ViT-Base contains aapproximately 886868686 86*illiorameters, several corably smaller archiectures have* chieved superior results. Conversely, models with comparable parameter counts often exhibited substantially different performance levels, suggesting that architectural design has a greater impact on synthetic-to-real transfer than network capacity itself.

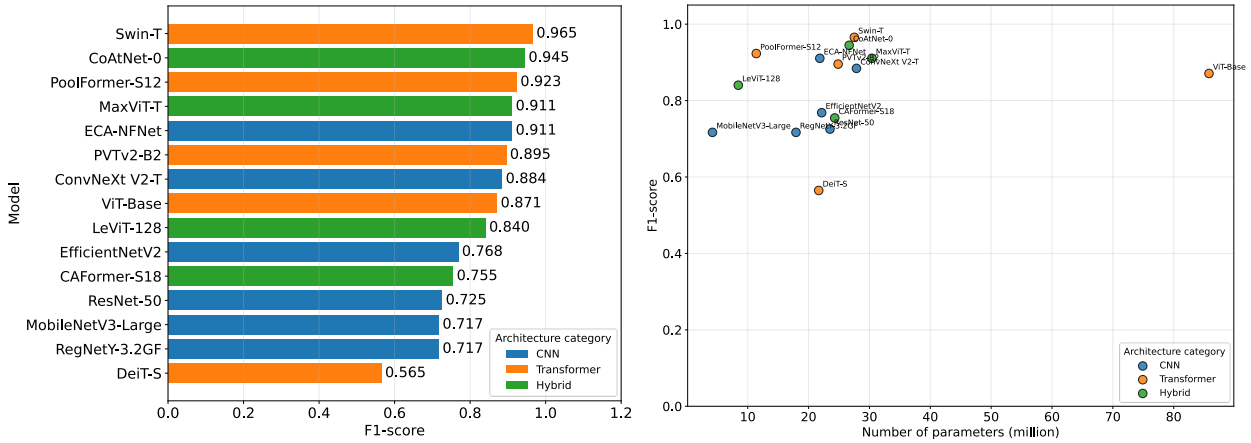


Figure 3. Overall comparison of all evaluated models.

Among the CNN-based models, ConvNeXt V2-T achieved the strongest performance, with ROC-AUC 0.982 and PR-AUC . 983.. Its ROC and Precision–Recall curves consistently dominate those of the remaining CNN architectures, indicating excellent discrimination capability across different operating thresholds. ResNet-50 and ECA-NFNet-L0 also demonstrated strong generalization performance, while RegNetY-3.2GF achieved slightly lower but still competitive results. The remaining CNN architectures exhibited considerably weaker performance. MobileNetV3-Large and EfficientNetV2 showed substantially lower ROC and Precision–Recall characteristics, indicating that these architectures struggle to preserve discriminative representations when transferred from synthetic to real-world images. Overall, the CNN family exhibited the largest performance variation among the evaluated architecture groups, demonstrating that convolutional architectures differ significantly in their ability to handle domain shifts caused by camera contamination.

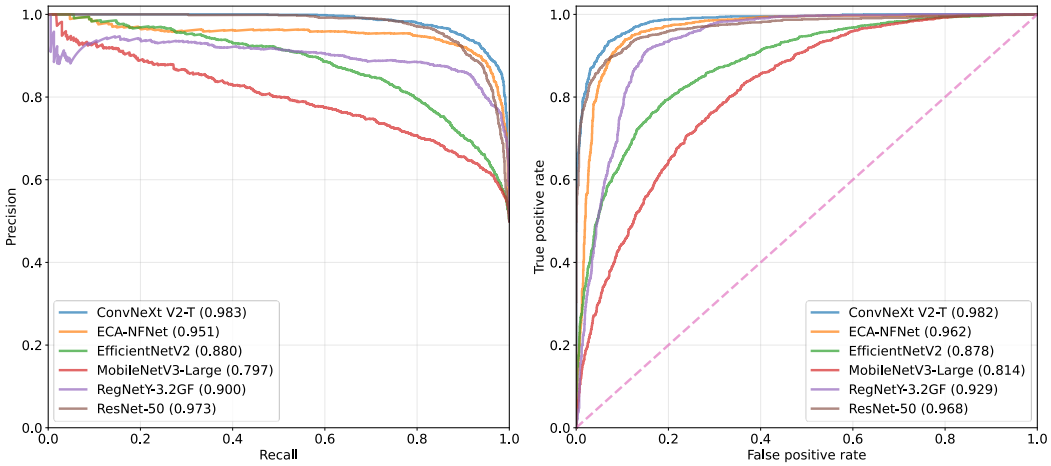


Figure 4. Precision-Recall (left) and ROC (right) curves for the CNN models.

The Transformer-based models generally demonstrated more consistent performance than the CNN architectures. Swin-T clearly outperformed all other Transformer models, achieving the highest ROC-AUC (0.996) and PR-AUC (0.996). Its Precision–Recall curve maintains very high precision even at high recall, indicating highly reliable discrimination between clean and contaminated images. ViT-Base also produced strong results despite being the largest evaluated model. However, its considerably larger number of trainable parameters did not translate into superior classification performance. PoolFormer-S12 and PVTv2-B2 exhibited similar operating characteristics, while DeiT-S consistently remained the weakest Transformer architecture. These observations indicate that modern hierarchical Transformer designs provide substantially better cross-domain robustness than earlier vision Transformer architectures.

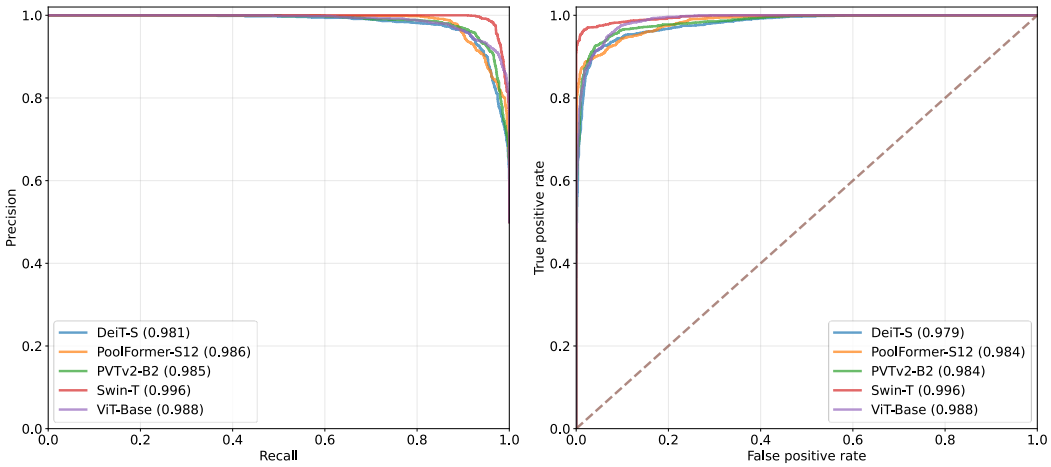


Figure 5. Precision-Recall (left) and ROC (right) curves for the transformer models.

The hybrid architectures combined the advantages of convolutional and Transformer-based feature extraction, resulting in consistently strong performance. In terms of both ROC-AUC and PR-AUC, MaxViT-T achieved the highest values among all evaluated models. CoAtNet-0 and CAFormer-S18 also demonstrated high classification accuracy, while LeViT-128 achieved competitive performance despite containing significantly fewer trainable parameters. Compared with CNN architectures, the hybrid models exhibited considerably smaller performance differences, suggesting that combining local convolutional processing with global attention mechanisms improves robustness to the synthetic-to-real domain shift.

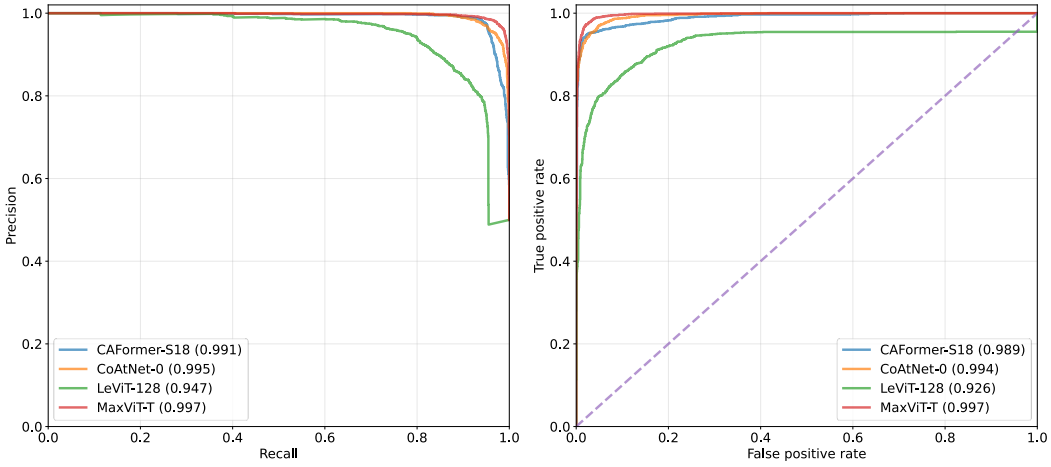


Figure 6. Precision-Recall (left) and ROC (right) curves for the hybrid models.

The confusion matrices provide further insight into the classification behavior of the evaluated models. The best-performing architectures, including Swin-T, CoAtNet-0, and PoolFormer-S12, produce only a small number of false-positive and false-negative predictions, indicating balanced performance across both classes. In contrast, lower-performing models exhibit substantially higher error rates. For example, DeiT-S frequently misclassifies both clean and contaminated images, confirming the observations from the ROC and Precision–Recall curves. Similar behavior is observed for ResNet-50 and MobileNetV3-Large. None of the evaluated architectures

exhibits a pronounced bias toward either class. Instead, performance differences primarily stem from the overall number of misclassified samples, suggesting that the balanced synthetic training dataset effectively mitigated class imbalance, while the remaining errors are mainly due to the domain gap. Overall, the results demonstrate that synthetic aiSim images provide an effective basis for camera contamination classification on real-world data. However, performance strongly depends on the network architecture, with hierarchical Transformer and hybrid models consistently outperforming conventional CNNs. In contrast, model complexity alone shows only a weak relationship with cross-domain generalization.

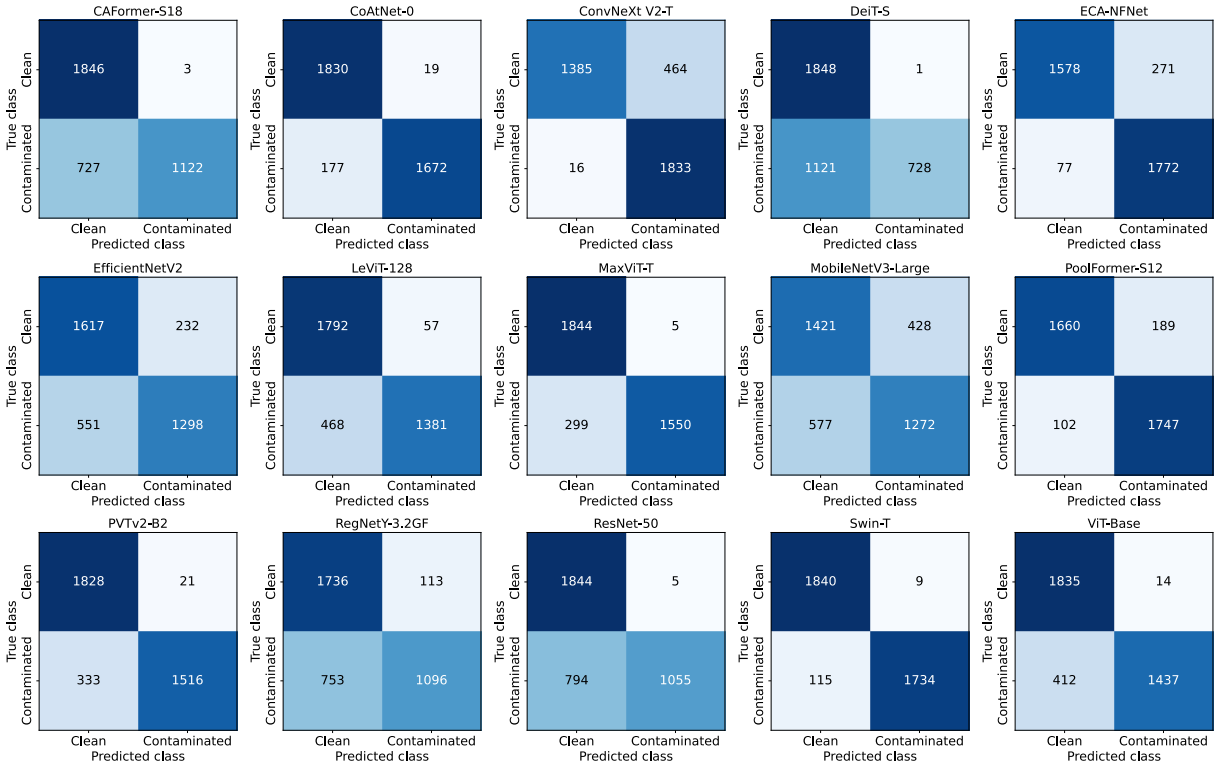


Figure 7. Confusion matrices of the evaluated models.

VI. CONCLUSIONS AND FUTURE WORK

This paper investigated the use of synthetic data for binary camera-contamination classification by evaluating the cross-domain generalization of 15 state-of-the-art CNN, transformer, and hybrid architectures. All models were trained exclusively on a synthetic aiSim dataset and evaluated on the real-world WoodScape dataset. The results demonstrate that synthetic training data can effectively support real-world camera contamination classification despite the domain gap between simulation and reality. Swin-T achieved the highest F1-score and ranked second in both ROC-AUC and PR-AUC, demonstrating consistently strong performance across all evaluation metrics. MaxViT-T achieved the highest ROC-AUC and PR-AUC values, indicating superior class-separation capability across different decision thresholds. Although no single architecture was superior across all metrics, Swin-T delivered the most consistently high performance while requiring fewer parameters than MaxViT-T. The experiments also showed that network architecture has a greater impact on generalization than model complexity, as several medium-sized models outperformed substantially larger ones. Overall, the findings indicate that synthetic datasets provide a practical alternative to expensive manual data collection for camera contamination classification, while emphasizing the importance of architecture selection for robust cross-domain

performance. Future work will validate the proposed approach using our own real-world measurements acquired under different driving conditions. In addition, the current binary classification task will be extended to other naturally occurring camera contaminations to investigate generalization across multiple contamination types. The long-term goal is to develop a unified camera contamination detection framework for improving perception monitoring and the reliability of autonomous driving systems under challenging environmental conditions.

ACKNOWLEDGEMENT

Supported by the EKÖP-25-4-II-SZE-142 University Research Fellowship Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

REFERENCES

[1] Zhang, Yuxiao, Alexander Carballo, Hanting Yang, and Kazuya Takeda. "Perception and Sensing for Autonomous Vehicles under Adverse Weather Conditions: A Survey." ISPRS Journal of Photogrammetry and Remote Sensing 196 (February 2023): 146–77. https://doi.org/10.1016/j.isprsjprs.2022.12.021.

[2] Sezgin, Fatih, Daniel Vriesman, Dagmar Steinhauser, Robert Lugner, and Thomas Brandmeier. "Safe Autonomous Driving in Adverse Weather: Sensor Evaluation and Performance Monitoring." 2023 IEEE Intelligent Vehicles Symposium (IV), June 4, 2023, 1–6. https://doi.org/10.1109/IV55152.2023.10186596.

[3] Kim, Hyunwoo, Yoseph Yang, Youngkwang Kim, Dong-Won Jang, Dongil Choi, Kang Park, Sangkug Chung, and Daegeun Kim. "Effect of Droplet Contamination on Camera Lens Surfaces: Degradation of Image Quality and Object Detection Performance." Applied Sciences 15, no. 5 (2025): 2690. https://doi.org/10.3390/app15052690.

[4] Csikor, Daniel, János Hollósi, and Ferenc Szauter. "Investigating the Impact of Environmental Factors on Autonomous Vehicle Sensors." SMTS 2025, October 28, 2025, 15. https://doi.org/10.3390/engproc2025113015.

[5] Csikor, Dániel, and János Hollósi. "Performance Degradation of Object Detection Neural Networks Under Natural Visual Contamination in Autonomous Driving." Computers 15, no. 4 (2026): 254. https://doi.org/10.3390/computers15040254.

[6] He, Da, Xiaoyu Shang, and Jiajia Luo. "Adherent Mist and Raindrop Removal from a Single Image Using Attentive Convolutional Network." Neurocomputing 505 (September 2022): 178–87. https://doi.org/10.1016/j.neucom.2022.07.032.

[7] Uricar, Michal, Pavel Krizek, Ganesh Sistu, and Senthil Yogamani. "SoilingNet: Soiling Detection on Automotive Surround-View Cameras." 2019 IEEE Intelligent Transportation Systems Conference (ITSC), October 2019, 67–72. https://doi.org/10.1109/ITSC.2019.8917178.

[8] Uricar, Michal, Ganesh Sistu, Hazem Rashed, et al. "Let's Get Dirty: GAN-Based Data Augmentation for Camera Lens Soiling Detection in Autonomous Driving." 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), January 2021, 766–75. https://doi.org/10.1109/WACV48630.2021.00081.

[9] Das, Arindam, Pavel Krizek, Ganesh Sistu, et al. "TiledSoilingNet: Tile-Level Soiling Detection on Automotive Surround-View Cameras Using Coverage Metric." 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), September 20, 2020, 1–6. https://doi.org/10.1109/ITSC45102.2020.9294677.

[10] Das, Arindam. "SoildNet: Soiling Degradation Detection in Autonomous Driving." arXiv:1911.01054. Preprint, arXiv, November 21, 2019. https://doi.org/10.48550/arXiv.1911.01054.

[11] Kim, Youngkwang, Woochan Kim, Jungwoo Yoon, Sangkug Chung, and Daegeun Kim. "Deep Learning-Based Multiple Droplet Contamination Detector for Vision Systems Using a You Only Look Once Algorithm." Information 15, no. 3 (2024): 134. https://doi.org/10.3390/info15030134.

[12] Soboleva, Vera, and Oleg Shipitko. "Raindrops on Windshield: Dataset and Lightweight Gradient-Based Detection Algorithm." 2021 IEEE Symposium Series on Computational Intelligence (SSCI), December 5, 2021, 1–7. https://doi.org/10.1109/SSCI50451.2021.9659915.

[13] Silva, Manuel, Antonio Seoane, Omar A. Mures, Antonio M. López y José A. Iglesias-Guitian. "Exploring the Effects of Synthetic Data Generation: A Case Study on Autonomous Driving for Semantic Segmentation." The Visual Computer 41, no. 10 (2025): 7379–97. https://doi.org/10.1007/s00371-025-03811-1.

[14] Miller, Leanne, Pedro J. Navarro, and Francisca Rosique. "EdgeNet: An End-to-End Deep Neural Network Pretrained with Synthetic Data for a Real-World Autonomous Driving Application." Sensors 25, no. 1 (2024): 89. https://doi.org/10.3390/s25010089.

[15] Xu, Wei, Dominique Gruyer, Alexandra Duminil, and Sio-Song Ieng. "Scenario-Driven Synthetic Data Generation Framework for Visual Perception Evaluation Under Adverse Driving Conditions." Sensors 26, no. 11 (2026): 3464. https://doi.org/10.3390/s26113464.

[16] Yogamani, Senthil, Ciaran Hughes, Jonathan Horgan, et al. "WoodScape: A Multi-Task, Multi-Camera Fisheye Dataset for Autonomous Driving." arXiv:1905.01489. Preprint, arXiv, July 2, 2021. https://doi.org/10.48550/arXiv.1905.01489.

[17] Yu, Weihao, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. "MetaFormer Baselines for Vision." IEEE Transactions on Pattern Analysis and Machine Intelligence 46, no. 2 (2024): 896–912. https://doi.org/10.1109/TPAMI.2023.3329173.

[18] Dai, Zihang, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. "CoAtNet: Marrying Convolution and Attention for All Data Sizes." Proceedings of the 35th International Conference on Neural Information Processing Systems (Red Hook, NY, USA), NIPS '21, 2021.

[19] Woo, Sanghyun, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. "ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders." 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, 16133–42. https://doi.org/10.1109/CVPR52729.2023.01548.

[20] Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jegou. "Training Data-Efficient Image Transformers & Distillation through Attention." In Proceedings of the 38th International Conference on Machine Learning, vol. 139, edited by Marina Meila and Tong Zhang. Proceedings of Machine Learning Research. PMLR, 2021.

[21] Wang, Qilong, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020, 11531–39. https://doi.org/10.1109/CVPR42600.2020.01155.

[22] Tan, Mingxing, and Quoc V. Le. "EfficientNetV2: Smaller Models and Faster Training." arXiv:2104.00298. Preprint, arXiv, June 23, 2021. https://doi.org/10.48550/arXiv.2104.00298.

[23] Graham, Ben, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. "LeViT: A Vision Transformer in ConvNet's Clothing for Faster Inference." 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, 12239–49. https://doi.org/10.1109/ICCV48922.2021.01204.

[24] Tu, Zhengzhong, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. "MaxViT: Multi-Axis Vision Transformer." In Computer Vision – ECCV 2022, vol. 13684, edited by Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner. Springer Nature Switzerland, 2022. https://doi.org/10.1007/978-3-031-20053-3_27.

[25] Howard, Andrew, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, Quoc Le. "Searching for MobileNetV3."

2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 2019, 1314–24. https://doi.org/10.1109/ICCV.2019.00140.

[26] Yu, Weihao, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. "MetaFormer Is Actually What You Need for Vision." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, 10809–19. https://doi.org/10.1109/CVPR52688.2022.01055.

[27] Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "PVT v2: Improved Baselines with Pyramid Vision Transformer." Computational Visual Media 8, no. 3 (2022): 415–24. https://doi.org/10.1007/s41095-022-0274-8.

[28] Radosavovic, Ilija, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. "Designing Network Design Spaces." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020, 10425–33. https://doi.org/10.1109/CVPR42600.2020.01044.

[29] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, 770–78. https://doi.org/10.1109/CVPR.2016.90.

[30] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

[31] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv:2010.11929. Preprint, arXiv, June 3, 2021. https://doi.org/10.48550/arXiv.2010.11929.

[32] Loshchilov, Ilya, and Frank Hutter. "Decoupled Weight Decay Regularization." arXiv:1711.05101. Preprint, arXiv, January 4, 2019. https://doi.org/10.48550/arXiv.1711.05101.

[33] Paszke, Adam, Sam Gross, Francisco Massa, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." arXiv.Org, December 3, 2019. https://arxiv.org/abs/1912.01703v1.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Evaluating TOE Determinants of Big Data Analytics Readiness through CRITIC–TOPSIS: Evidence from Albania

Megi Marku
Department of Mathematics and Informatics
Agricultural University of Tirana
Tirana, Albania
mmarku@ubt.edu.al
ORCID: 0009-0004-1575-2990

Abstract

As organizations continue to invest in Big Data Analytics (BDA), understanding the conditions that support effective implementation has become increasingly important. Firms sharing the same adoption status often differ substantially in their technological, organizational, and environmental preparedness. This study applies a Multi-Criteria Decision-Making (MCDM) approach, grounded in the Technology–Organization–Environment (TOE) framework, to evaluate post-adoption readiness among BDA-adopting firms. The analysis draws on survey data collected from Albanian large taxpayers actively using BDA between May 2024 and May 2025. Nine TOE determinants served as evaluation criteria. The CRITIC method derived objective criterion weights, while TOPSIS computed readiness scores, ranking firms by proximity to ideal conditions. Risk (0.185) and Complexity (0.176) exhibited the greatest discriminating power, together accounting for roughly 36% of the total weight, followed by Organizational Readiness (0.115); External Support received the lowest weight. Readiness scores ranged from 0.424 to 0.691 (mean: 0.529), indicating considerable variation among adopters despite their shared adoption status. A sensitivity analysis against equal weighting (Spearman $\rho = 0.36$) confirmed that objective weighting materially affected prioritization. The study extends the TOE framework toward a post-adoption readiness perspective and demonstrates the applicability of CRITIC–TOPSIS for evaluating organizational preparedness in an emerging-economy context.

Keywords: Big Data Analytics, TOE framework, CRITIC, TOPSIS, MCDM

I. INTRODUCTION

BDA has become an important capability for organizations seeking to improve decision-making, operational efficiency, competitiveness, and business performance [1], [2]. As firms increasingly rely on large volumes of structured and unstructured data, the ability to transform data into actionable insights is no longer only a technological issue, but also an organizational and strategic requirement [3]. However, adopting BDA does not necessarily imply that all firms are equally prepared to use it effectively. Organizations that report BDA adoption may still differ considerably in terms of technological infrastructure, organizational capabilities, managerial support, external pressures, and perceived implementation barriers. Most previous studies on BDA have focused on explaining adoption decisions by comparing adopters and non-adopters [4]. While this perspective is important, it provides limited insight into the readiness levels of firms that have already adopted BDA. Post-adoption readiness deserves particular attention because the effectiveness of BDA depends not only on whether the technology is adopted, but also on the conditions under which it is implemented and used. Firms may formally adopt BDA while still facing problems related to complexity, risk, lack of organizational readiness, insufficient external support, or limited managerial commitment [5]. To address this issue, this study evaluates BDA readiness among firms that have already adopted BDA, using the TOE framework as the theoretical basis.

The TOE framework is appropriate for organizational-level technology studies because it explains adoption and implementation through three interrelated contexts: technological, organizational, and environmental [6]. In this study, nine TOE determinants are considered: Relative Advantage, Compatibility, Complexity, Risk, Top Management Support, Organizational Readiness, Competitive Pressure, External Support, and Government Support. Methodologically, the study employs an MCDM approach, combining the CRITIC and TOPSIS methods. The CRITIC method is used to determine objective criterion weights based on the variability and interrelationships among criteria. In contrast, TOPSIS is used to rank firms by their proximity to the ideal TOE readiness conditions. This integrated approach enables identification of which TOE determinants contribute most strongly to readiness and ranking of BDA-adopting firms by overall preparedness. The contribution of this study is threefold. First, it extends BDA adoption research by shifting the focus from adoption versus non-adoption to post-adoption readiness. Second, it introduces a CRITIC–TOPSIS framework for evaluating readiness among BDA adopters. Third, it provides empirical evidence from large Albanian taxpayers, offering insights in an emerging-economy context where digital transformation and advanced analytics capabilities are still developing. Accordingly, the study addresses two research questions:

RQ1: Which TOE determinants contribute most strongly to differentiating readiness levels among firms that have already adopted BDA?

RQ2: To what extent do BDA-adopting firms differ in their overall post-adoption readiness?

II. LITERATURE REVIEW

The TOE framework has been widely used to explain technology adoption and implementation at the organizational level. Originally introduced by [6], the framework suggests that technological innovation is shaped by three main contexts: technological, organizational, and environmental. Unlike individual-level acceptance models, TOE is particularly well-suited to firm-level analysis because it captures both internal and external factors that influence the adoption and use of information technologies. In the context of BDA, the technological dimension includes factors such as relative advantage, compatibility, complexity, and risk. Relative advantage refers to the perceived benefits that BDA can offer compared with existing analytical tools or decision-making practices. Compatibility reflects the extent to which BDA fits with existing systems, business processes, and organizational needs. Complexity, on the other hand, represents the perceived difficulty of implementing and using BDA technologies. Risk is also an important technological barrier, especially because BDA involves large-scale data processing, privacy concerns, security issues, and uncertainty regarding implementation outcomes. Prior studies emphasize that technological challenges, including fragmented data systems, integration issues, data security concerns, and system complexity, can significantly affect BDA implementation [2], [5]. The organizational dimension includes internal capabilities and resources that support successful implementation. Top management support is frequently identified as a key factor in technology adoption because leadership commitment influences resource allocation, strategic prioritization, and employee engagement. Organizational readiness is also essential, as firms need adequate financial, technical, and human resources to implement BDA effectively.

Studies on BDA adoption suggest that organizations with stronger data-driven cultures, better internal capabilities, and higher managerial commitment are more likely to implement analytics initiatives successfully [2], [3]. Consistent with this, technological and organizational factors are the more significant determinants of BDA adoption among firms [7]. The environmental dimension captures external pressures and support mechanisms that influence organizational decisions. Competitive pressure may encourage firms to adopt BDA to improve performance, respond to market changes, and remain competitive. External support, including vendor assistance, consulting services, and technology partners, may reduce implementation barriers. Government

support and regulatory conditions may also influence the adoption and use of data-driven technologies, particularly in emerging economies where institutional support can shape digital transformation strategies. Previous studies indicate that external competition, regulatory requirements, and institutional support can affect BDA adoption and implementation outcomes [5]. Although the TOE framework provides a comprehensive explanation of BDA adoption, most studies have focused on identifying determinants of adoption rather than evaluating readiness among existing adopters. This represents an important gap because adoption status alone does not indicate whether a firm is fully prepared to benefit from BDA. Firms may adopt BDA but differ substantially in their technological maturity, organizational capacity, and environmental support. Therefore, readiness assessment requires a more comparative and multi-dimensional approach. Multi-Criteria Decision-Making methods are useful for this purpose because they allow researchers to evaluate alternatives based on multiple criteria simultaneously. The CRITIC method is an objective weighting technique that determines criterion importance based on each criterion's variability and its conflict with other criteria. Criteria that exhibit greater variation and provide more unique information receive greater weight. This makes CRITIC suitable for organizational datasets in which subjective, expert-based weighting may introduce bias [8], [9]. TOPSIS is another widely used MCDM method for ranking alternatives based on their distances from the ideal and negative ideal solutions. The method assumes that the best alternative should be closest to the positive ideal solution and farthest from the negative ideal solution. In information systems and project selection contexts, TOPSIS has been applied to rank alternatives according to criteria such as reliability, cost-effectiveness, performance, and implementation feasibility [9], [10]. The integration of CRITIC and TOPSIS provides a structured, data-driven framework for readiness evaluation and has been applied effectively in recent multi-criteria evaluation studies [11]. Recent TOE-based studies have addressed BDA adoption in SMEs [12] and pre-adoption readiness in the health sector [13]; however, post-adoption readiness among firms that have already implemented BDA remains comparatively underexplored. CRITIC determines the objective importance of TOE determinants, while TOPSIS transforms these weighted criteria into readiness scores and rankings. This combined approach is especially appropriate for the present study because BDA readiness is multidimensional and cannot be captured by a single indicator. By applying CRITIC–TOPSIS to BDA-adopting firms, this study contributes to the literature by offering a post-adoption readiness assessment framework that identifies both the most influential TOE determinants and firms' relative preparedness.

To position the present study within the existing literature, Table 1 summarises representative TOE-based studies on BDA adoption and readiness, highlighting their contexts, methods, and analytical foci.

Table 1. Representative TOE-based studies on BDA adoption and readiness

Study	Context	Method	Focus
Maroufkhani et al. [2]	SMEs, Malaysia	PLS-SEM	Adoption determinants and performance
Mikalef et al. [3]	Firms, mixed-method	SEM / qualitative	BDA capability and firm performance
Al-shanableh et al. [4]	SMEs, Jordan	PLS-SEM	Adoption (extended TOE + DOI)
Kumar et al. [12]	SMEs, India	PLS-SEM	Adoption and market/financial performance
Assaye et al. [13]	Health sector, Ethiopia	SEM	Readiness (pre-adoption)

III. METHODOLOGY

This study applies an MCDM approach to evaluate the readiness of firms that have adopted BDA. Data were collected between May 2024 and May 2025 through a structured questionnaire distributed to large taxpayers. Because the survey employed a two-instrument design, missing values were addressed using the Expectation–Maximization (EM) algorithm, which preserves the aggregate means and covariance structure on which the CRITIC–TOPSIS criteria are constructed. As the analysis relies on construct-level scores derived from these aggregate statistics rather than on individual-level responses, the resulting readiness ranking is robust to the specific imputed values. The EM procedure was applied to the full sample of 121 large taxpayers, following the TOE measurement structure, after which the CRITIC–TOPSIS method was applied to firms reporting active BDA adoption. The study focuses not on explaining adoption versus non-adoption, but on assessing the relative readiness levels of firms that have already implemented BDA. The analytical framework is based on the TOE model. Nine determinants were used as evaluation criteria: Relative Advantage (RA), Compatibility (CO), Complexity (C), Risk (RI), Top Management Support (M), Organizational Readiness (OR), Competitive Pressure (CP), External Support (ES), and Government Support (G). Although the TOE determinants were originally formulated in the context of the adoption decision, the underlying constructs remain conceptually relevant during the implementation stage. Complexity, risk, organizational readiness, and related conditions do not cease to operate once a technology is formally adopted; rather, they continue to shape how effectively firms integrate, manage, and sustain BDA in practice. The determinants are therefore interpreted as indicators of ongoing post-adoption readiness rather than of the initial adoption choice. Each construct was measured through multiple questionnaire items using a five-point Likert scale ranging from strong disagreement to strong agreement. Before the MCDM analysis, the psychometric properties of the measurement instrument were assessed. Internal consistency was evaluated using Cronbach's Alpha, while construct reliability was assessed using composite reliability (CR). Convergent validity was assessed using the Average Variance Extracted (AVE), and discriminant validity was confirmed following the Fornell–Larcker criterion, as the square root of each construct's AVE exceeded its correlations with the remaining constructs. The results confirmed satisfactory reliability and validity: all constructs exceeded the recommended thresholds, with Cronbach's alpha ranging from 0.845 to 0.962, composite reliability (CR) from 0.851 to 0.963, and AVE from 0.657 to 0.843.

Composite scores for each TOE determinant were calculated by averaging the corresponding questionnaire items. Because each construct was first validated as a reliable and unidimensional measure — with Cronbach's alpha, composite reliability, and AVE all exceeding recommended thresholds and discriminant validity confirmed through the Fornell–Larcker criterion — the averaging of items into a single construct score constitutes a justified representation of the underlying latent variable. This aggregation is consistent with standard practice when MCDM techniques are applied to validated survey constructs. These scores were then used as evaluation criteria in the MCDM framework. The recommended thresholds applied were 0.70 for Cronbach's alpha and composite reliability, and 0.50 for AVE [14]. To determine the relative importance of the TOE determinants, the CRITIC method was employed. CRITIC is an objective weighting technique that derives criterion weights directly from the observed data rather than relying on expert judgments. The method simultaneously accounts for the variability of each criterion and its degree of correlation with the other criteria. Consequently, criteria exhibiting higher variability and providing more unique information receive greater weights. This characteristic makes CRITIC particularly suitable for organizational datasets where subjective weighting may introduce bias. It should be emphasized that CRITIC weights express the discriminating and informational power of each criterion within the observed sample rather than its intrinsic strategic or managerial importance. A high weight, therefore, indicates that firms differ substantially on that criterion, making it informative for distinguishing readiness levels, which is precisely the analytical objective of this study, and not that the criterion is theoretically more important than the others.

Following the weighting stage, the TOPSIS was applied to assess the readiness level of each BDA-adopting firm. TOPSIS is one of the most widely used MCDM techniques because it evaluates alternatives based on their proximity to the ideal solution and their distance from the negative ideal solution. In the present study, Relative Advantage, Compatibility, Top Management Support, Organizational Readiness, Competitive Pressure, External Support, and Government Support were treated as benefit criteria. In contrast, Complexity and Risk were treated as cost criteria, as higher values represent barriers to successful BDA implementation. The decision matrix was normalized using vector normalization before applying the criterion weights. The combined CRITIC-TOPSIS framework generated two principal outputs. First, it produced objective weights indicating the relative importance of each TOE determinant. Second, it generated a readiness score for each firm, enabling ranking of BDA adopters by overall preparedness. This approach provides a comprehensive, data-driven assessment of organizational readiness and enables identification of the TOE factors that contribute most strongly to successful BDA implementation. All analyses were conducted in R using customized scripts for data preparation, CRITIC weighting, and TOPSIS ranking procedures. The overall analytical workflow, from data collection to sensitivity analysis, is summarised in Figure 1.



Figure 1. CRITIC–TOPSIS methodological workflow

IV. RESULTS

The CRITIC-TOPSIS framework was applied to assess the readiness levels of BDA-adopting firms and to identify the relative importance of TOE determinants. The analysis was conducted in two stages.
First, the CRITIC method was used to derive objective criterion weights based on the variability and informational content of the TOE factors. Second, the TOPSIS method was employed to calculate readiness scores and rank firms according to their proximity to the ideal TOE conditions. The results provide insights into the determinants that contribute most strongly to organizational readiness and reveal differences in readiness levels among firms that have already implemented BDA.

Table 2 presents the objective weights generated through the CRITIC method. The results indicate that Risk (0.185) and Complexity (0.176) received the highest weights among all TOE determinants. Together, these two determinants account for 36.15% of the total weight, emphasizing the dominant role of technological barriers in differentiating readiness levels among BDA adopters.
This suggests that these two factors provide the greatest information and discrimination among BDA-adopting firms. In other words, firms differ substantially in their perceptions of implementation risk and technological complexity, making these determinants particularly influential in evaluating organizational readiness. Organizational Readiness ranked third with a weight of 0.115, followed by Relative Advantage (0.105). These findings highlight the importance of internal capabilities and the perceived benefits of BDA adoption. In contrast, External Support (0.075), Competitive Pressure (0.079), and Top Management Support (0.082) received comparatively lower weights, indicating lower variability and lower informational contribution within the sample.

Table 2. CRITIC Weights of TOE Determinants

Rank	Criterion	Weight
1	Risk (RI)	0.185
2	Complexity (C)	0.176
3	Organizational Readiness (OR)	0.115
4	Relative Advantage (RA)	0.105
5	Compatibility (CO)	0.091
6	Government Support (G)	0.091
7	Top Management Support (M)	0.082
8	Competitive Pressure (CP)	0.079
9	External Support (ES)	0.075

Figure 2 provides a visual representation of these weights, ordered from highest to lowest, making the pronounced gap between the two technological criteria and the remaining determinants readily apparent.

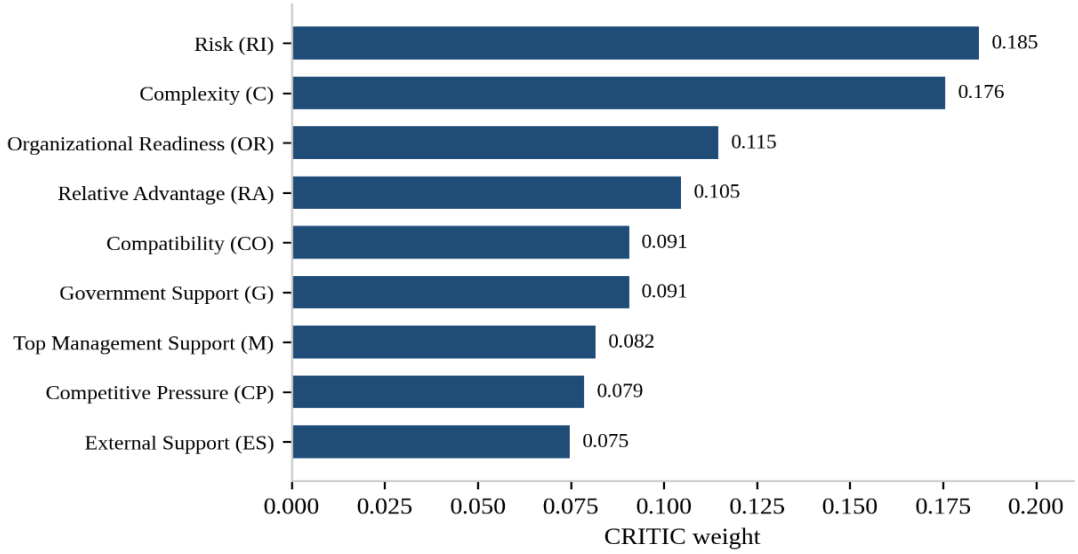


Figure 2. CRITIC weights of the nine TOE determinants

Table 3 presents the aggregated importance of the three TOE dimensions based on the CRITIC-derived weights of their respective determinants. The technological dimension accounts for more than half of the total importance (0.557), indicating that technological considerations dominate the readiness assessment of BDA-adopting firms. This result is primarily driven by the high weights assigned to Risk and Complexity, which emerged as the two most influential determinants in the analysis. The environmental dimension ranks second with a cumulative weight of 0.245, suggesting that external conditions, including competitive pressure, external support, and government support, contribute to organizational readiness but play a less prominent role than technological factors. The organizational dimension received the lowest overall weight (0.197). Although Organizational Readiness emerged as one of the most important individual determinants, the combined contribution of organizational factors remains lower than that of the technological dimension. It should be noted, however, that the aggregated dimensional weights are partly a function of the number of criteria assigned

to each dimension. The technological dimension comprises four determinants, the environmental dimension three, and the organizational dimension only two. When the average weight per criterion is considered, the gap narrows: 0.139 for the technological dimension, 0.099 for the organizational dimension, and 0.082 for the environmental dimension. The technological dimension, therefore, remains dominant on both measures, but the comparison of raw dimensional sums should be interpreted with this structural asymmetry in consideration. The findings indicate that technological conditions represent the primary source of differentiation among BDA-adopting firms and therefore deserve particular attention when assessing organizational readiness.

Table 3. Relative Importance of TOE Dimensions Based on CRITIC Weights

TOE Dimension	Weight	Rank
Technological	0.557	1
Environmental	0.245	2
Organizational	0.197	3

Table 4 summarizes the distribution of TOE-TOPSIS readiness scores across the BDA-adopting firms. The readiness scores range from 0.424 to 0.691, with an average value of 0.529. The relatively wide range of scores indicates that not all adopters share the same level of readiness, even after implementing BDA solutions. The median score of 0.526 indicates that approximately half of the firms exhibit readiness levels above this threshold, while the remaining firms display lower preparedness. The difference between the minimum and maximum values further suggests considerable heterogeneity among adopters regarding their technological, organizational, and environmental conditions.

To assess the robustness of the results to the choice of weighting method, a sensitivity analysis was conducted comparing the CRITIC-weighted ranking with an equal-weight TOPSIS solution. The Spearman rank correlation between the two rankings was $\rho = 0.36$, indicating only a weak-to-moderate association. The choice of weighting scheme, therefore, materially affected the resulting prioritization of firms.

Table 4. Descriptive Statistics of TOE-TOPSIS Scores

Statistic	Value
Minimum	0.424
1st Quartile	0.476
Median	0.526
Mean	0.529
3rd Quartile	0.585
Maximum	0.691
St. deviation	0.065

V. DISCUSSION

This study examined post-adoption BDA readiness among Albanian large taxpayers by integrating the TOE framework with the CRITIC–TOPSIS methodology. The findings demonstrate that firms reporting active BDA adoption remain substantially heterogeneous in their technological, organizational, and environmental preparedness. This supports the central argument that adoption should not be treated as a binary organizational condition: firms may formally implement BDA technologies while differing considerably in their capacity to manage implementation barriers, integrate analytical systems, and mobilize internal resources. Risk and

Complexity emerged as the criteria with the greatest discriminating power, together accounting for more than one-third of the total weight. This is in line with research emphasizing that data security concerns, implementation uncertainty, and technical integration challenges remain central obstacles even after adoption, consistent with readiness assessments in other sectors that emphasize organizational preparedness [13]. Maroufkhani et al. [2] identify complexity, uncertainty, and insecurity among the significant determinants of BDA adoption, while Hamed and Bohari [5] similarly emphasize technological and infrastructural barriers to analytics adoption in developing-economy contexts. The high weight on Risk indicates that adopting firms differ considerably in their perceptions of data-related uncertainty, privacy, and security — a relevant finding given that BDA systems typically integrate large, heterogeneous datasets across multiple organizational functions. Readiness, therefore, cannot be evaluated solely by the presence of analytical technologies; it must also capture the firm's capacity to identify and mitigate implementation risks.

Complexity, ranked second, indicates that the perceived difficulty of implementing and integrating BDA is an important source of variation among adopters. Although all firms in the sample reported adoption, they differ in system sophistication, availability of technical expertise, compatibility of data sources, and the extent to which analytics are embedded in organizational routines. Adoption does not necessarily imply successful integration: firms may implement BDA tools at a formal or limited level while continuing to face difficulties with technical skills, interoperability, and process redesign. The prominence of these two criteria explains why the technological dimension received the highest aggregated weight, exceeding the environmental and organizational dimensions. This should not be read as evidence that the latter are unimportant. CRITIC weights reflect variability and informational contribution within the observed sample rather than causal effects, so the technological dimension appears dominant because technological conditions differentiate firms more strongly than other conditions in this particular group of adopters. Organizational Readiness ranked third, confirming the importance of internal resources and implementation capacity. Maroufkhani et al. [2] identify organizational readiness and top management support as significant determinants of BDA adoption, while Mikalef et al. [3] emphasize that organizational capabilities and data-driven culture underpin the development of analytics capacity. Even where technological barriers dominate the readiness assessment, internal capacity remains a foundation for sustaining implementation. The weight assigned to Relative Advantage further suggests that perceived benefits remain relevant after adoption. Traditionally examined as a predictor of the adoption decision, perceived usefulness may also influence whether firms deepen their analytical capabilities or limit implementation. Top Management Support received a comparatively lower weight, which does not imply that managerial support is unimportant. Since all sampled organizations had already implemented BDA, a minimum level of managerial commitment was likely present across the sample, reducing variability and, therefore, the informational contribution. A similar logic applies to the environmental dimension, which ranked second in aggregate. Government Support may be especially relevant in emerging economies, where institutional policies and public investment shape firms' capacity to develop analytics capabilities, and Competitive Pressure may encourage faster analytical response to market developments. External Support received the lowest weight, possibly because large taxpayers command stronger internal technical and financial resources than smaller firms and depend less on external providers, or because vendor and consultancy support is relatively uniform across the sample. This diverges from contexts in which external expertise is considered essential for overcoming resource constraints, suggesting that its importance depends on firm size, market development, and the availability of specialized providers. The distribution of TOPSIS scores, ranging from 0.424 to 0.691, provides further evidence of heterogeneity: firms with identical formal adoption status occupy considerably different levels of preparedness. This moves the analysis beyond the conventional adopter/non-adopter distinction and indicates that adoption represents only one stage of a broader implementation process. The sensitivity analysis reinforces the methodological value of objective weighting: the weak-to-moderate Spearman correlation between the CRITIC-weighted and equal-weight rankings shows that the choice of weighting scheme materially affected firm prioritization, since equal weighting would assume identical contribution across all nine determinants, whereas CRITIC captured differences in variability and informational

content. These findings translate into concrete implications for practice. For managers, post-adoption improvement efforts should be directed first toward risk management and complexity reduction, since these dimensions account for the largest share of variation among adopters. Practical measures include strengthening data governance and security protocols, clarifying accountability for analytical outputs, investing in interoperability across systems, and simplifying the interface between analytical tools and existing business processes. The readiness scores themselves serve as a benchmarking instrument: firms positioned in the lower quartile of the distribution can identify the specific criteria on which they deviate most from the ideal solution and target interventions accordingly. For policymakers, the low weight assigned to External Support indicates that vendor and consultancy provision currently produces little differentiation among large Albanian enterprises. This may reflect either the internal resource strength of this segment or an underdeveloped market for specialized analytics services. Targeted programs aimed at smaller firms, where external expertise is more likely to be decisive, may therefore yield greater returns than broad support schemes. This aligns with evidence from SME contexts, where technological and organizational determinants strongly shape BDA adoption outcomes [12]. These contributions should nonetheless be interpreted in light of several limitations. First, the analysis is based on BDA-adopting firms drawn from the population of Albanian large taxpayers. At the same time, this group represents the segment of the economy most likely to implement advanced analytics; the findings cannot be generalized to small and medium-sized enterprises or to other national contexts. Second, the data are cross-sectional, which precludes any assessment of how readiness evolves or how it translates into subsequent performance outcomes. Third, all TOE determinants were measured using self-reported perceptual items administered to a single respondent per firm, introducing potential common method bias and social desirability effects; objective indicators of technological infrastructure and analytical capability would strengthen future assessments. Fourth, and central to the interpretation of the results, CRITIC weights are sample-dependent by construction: they reflect the variability and intercorrelation of criteria within this particular set of firms rather than any universal importance of the TOE determinants, so a different sample would likely produce different weights, and the reported values should be read as descriptive of this population rather than as generalizable parameters. Finally, the CRITIC–TOPSIS framework is comparative rather than explanatory: it ranks firms relative to ideal and anti-ideal conditions. Still, it does not estimate the causal effects of TOE determinants on implementation outcomes.

VI. CONCLUSIONS

This study demonstrates that BDA adoption is not a binary organizational state. Among firms reporting active adoption, readiness scores ranged from 0.424 to 0.691, indicating substantial heterogeneity in the technological, organizational, and environmental conditions that underpin implementation. Risk and Complexity were the determinants that most strongly differentiated firms, indicating that technological conditions rather than adoption status itself are the primary source of variation in post-adoption readiness within this population of adopters. Theoretically, the study extends the TOE framework from an adoption-oriented model toward a post-adoption readiness framework, demonstrating that the same three dimensions can assess variation among firms that have already adopted a technology. This matters because implementation effectiveness depends on the continuing alignment of technological infrastructure, organizational capabilities, and environmental conditions, not on the adoption decision alone. Methodologically, CRITIC–TOPSIS provides a structured procedure for deriving criterion weights, readiness scores, and rankings from multidimensional TOE information. Unlike regression-based approaches, it does not estimate causal effects; it provides a comparative evaluation tool that identifies the most informative criteria and benchmarks organizations against ideal readiness conditions, making it well-suited to managerial diagnosis. By operationalizing the TOE framework through CRITIC–TOPSIS, the study offers both a theoretical repositioning and a practical diagnostic tool for evaluating organizational preparedness in emerging-economy contexts. Future research could build on these results by

addressing the limitations noted above: applying the framework longitudinally to capture how readiness evolves and translates into performance, extending it to SMEs and to cross-country samples, validating readiness scores against objective performance measures, and comparing CRITIC–TOPSIS against alternative MCDM configurations such as entropy-weighted TOPSIS, standard-deviation weighting, or VIKOR to test whether the observed ranking is sensitive to the choice of weighting and aggregation method.

REFERENCES

[1] S. F. Wamba, A. Gunasekaran, S. Akter, S. J. Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," *J. Bus. Res.*, vol. 70, pp. 356–365, 2017.

[2] P. Maroufkhani, M.-L. Tseng, M. Iranmanesh, W. K. W. Ismail, and H. Khalid, "Big data analytics adoption: Determinants and performances among small to medium-sized enterprises," *Int. J. Inf. Manage*, vol. 54, art. 102190, 2020.

[3] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, "Big data analytics and firm performance: Findings from a mixed-method approach," *J. Bus. Res.*, vol. 98, pp. 261–276, 2019.

[4] N. Al-shanableh, M. Alzyoud, S. Alomar, Y. Kilani, E. Nashnush, S. Al-Hawary, and A. Al-Momani, "The adoption of big data analytics in Jordanian SMEs: An extended technology organization environment framework with diffusion of innovation and perceived usefulness," *Int. J. Data Netw. Sci.*, vol. 8, no. 2, pp. 753–764, 2024.

[5] A. Hamed and A. M. Bohari, "Adoption of big data analytics in medium-large supply chain firms in Saudi Arabia," *Knowl. Perform. Manage.*, vol. 6, no. 1, pp. 62–74, 2022.

[6] L. G. Tornatzky and M. Fleischer, *The Processes of Technological Innovation*. Lexington, MA: Lexington Books, 1990.

[7] P. Maroufkhani, W. K. Wan Ismail, and M. Ghobakhloo, "Big data analytics adoption model for small and medium enterprises," *J. Sci. Technol. Policy Manage.*, vol. 11, no. 4, pp. 483–513, 2020.

[8] D. Diakoulaki, G. Mavrotas, and L. Papayannakis, "Determining objective weights in multiple criteria problems: The CRITIC method," *Comput. Oper. Res.*, vol. 22, no. 7, pp. 763–770, 1995.

[9] J. Huang, "Combining entropy weight and TOPSIS method for information system selection," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, 2008, pp. 1281–1284.

[10] C.-L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications*. Berlin, Germany: Springer, 1981.

[11] M. Wei, J. Zheng, S. Zeng, and Y. Jin, "Evaluation of high quality and full employment based on CRITIC-entropy-TOPSIS multi-criteria framework," *Int. J. Intell. Comput. Cybern.*, vol. 17, no. 3, pp. 465–485, 2024.

[12] J. Kumar, G. Rani, M. Rani, and V. Rani, "Big data analytics adoption and its impact on SME market and financial performance: An analysis using the Technology–Organization–Environment (TOE) framework," *Creativity and Innovation Management*, vol. 34, no. 3, pp. 760–777, 2025.

[13] B. T. Assaye et al., "Readiness of big health data analytics by technology-organization-environment (TOE) framework in Ethiopian health sectors," *Heliyon*, vol. 10, no. 19, art. e38570, 2024.

[14] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *J. Mark—Res.*, vol. 18, no. 1, pp. 39–50, 1981.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Traffic Accident Severity Prediction with an Interpretable Crash Severity Risk Score

Sevcel Gözde BILGIN
Gazi University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering
Ankara, Türkiye
23833304001@gazi.edu.tr
ORCID:0000-0002-3786-7017

Cemal KOCAK
Gazi University, Faculty of Technology, Department of Computer Engineering
Ankara, Türkiye
ccckocak@gazi.edu.tr
ORCID:0000-0002-8902-0934

Abstract
This study proposes an interpretable and leakage-aware machine learning framework for classifying traffic accidents as fatal or injury-related. Vehicle, driver, and accident records from the General Directorate of Security were integrated, and an Ankara subset covering 2020–2025 was constructed. Because multiple vehicle records may correspond to the same accident, an accident-level stratified split based on accident year and KazaId was used to prevent cross-set overlap. Fatal outcomes accounted for approximately 1.45% of vehicle accident records, resulting in a highly imbalanced classification problem. The proposed Crash Severity Risk Score (CSRS) combines six risk dimensions—vehicle, driver, road, environment, time, and accident occurrence type—into a 0–12 score. To prevent target leakage, CSRS mappings were estimated exclusively from training data using out-of-fold estimation. On the held-out test set, mean CSRS was 5.996 for injury records and 7.718 for fatal records. Association Rule Mining was independently used to identify interpretable fatal-risk patterns. Six classifiers were evaluated with and without CSRS using validation-selected decision thresholds. LightGBM with CSRS achieved the highest held-out Average Precision (0.1462) and fatal-class F1-score (0.2198), while Balanced Random Forest with CSRS achieved the highest ROC-AUC (0.8522). Temporal and cross-province validation showed reduced performance under distribution shift, while the contribution of CSRS remained model-dependent. Overall, the results support CSRS as an interpretable and complementary feature for predicting traffic accident severity.

Keywords: Traffic Accident Severity, Fatal Accident Prediction, Crash Severity Risk Score, Association Rule Mining, Imbalanced Classification, LightGBM, XGBoost

I. INTRODUCTION

Traffic accidents are a critical road safety problem, not only because of their frequency but also because of the fatalities and injuries they cause. According to the World Health Organization, accurate analysis of accident severity has direct implications for public safety, healthcare systems, and the development of effective road safety strategies. Traffic accident severity prediction aims to estimate accident outcomes using historical characteristics related to the vehicle, driver, road, environment, time, and accident occurrence [1]. The limitations of traditional statistical methods in modeling large datasets and complex nonlinear relationships have led to the increasing use of Machine Learning (ML) and Deep Learning (DL) methods in this field [2]. These predictive approaches can support the identification of high-risk conditions, the development of sustainable road safety policies, and the development of data-driven decision support systems. The main motivation of this study is that fatal traffic accidents constitute a very small minority class in the dataset, making conventional accuracy-oriented evaluations potentially misleading [3], [4], [5]. For instance, when fatal records account for only approximately 1.45% of the held-out test observations, a classifier may achieve high overall accuracy while detecting only a limited proportion of fatal cases. Therefore, model performance in this study is evaluated using

minority-class-oriented measures, including precision, recall, F1-score, ROC-AUC, and Average Precision (AP), together with decision-threshold analysis. Prior studies have emphasized feature selection, deep learning, and hybrid architectures, whereas the development and independent evaluation of interpretable composite risk features remain comparatively limited [6], [7]. To address this gap, this study proposes an interpretable framework that combines Association Rule Mining (ARM), the Crash Severity Risk Score (CSRS), and machine learning-based severity classification. ARM is employed as an independent explanatory analysis to identify combinations of vehicle-, driver-, road-, environmental-, temporal-, and accident-occurrence-related conditions associated with fatal outcomes. Separately, CSRS summarizes six corresponding risk dimensions into a single numerical feature ranging from 0 to 12. To minimize target leakage, target-informed CSRS mappings are learned exclusively from training data, and out-of-fold score generation is used for training observations. The incremental predictive contribution of CSRS is then evaluated by comparing otherwise identical machine learning models with and without the score. In addition to an accident-level held-out test evaluation, temporal and geographic validation, paired bootstrap analysis, and CSRS sensitivity analysis were used to assess the robustness and generalizability of the proposed framework.

II. RELATED WORKS

Traffic accident severity prediction has been widely investigated using statistical, machine learning, and deep learning approaches. Systematic reviews have shown an increasing use of data-driven methods in road-safety analysis because these methods can capture complex and nonlinear relationships among driver, vehicle, road, environmental, and crash-related factors [2], [8]. Previous studies have employed feature-selection-based machine learning models, deep neural networks, and hybrid machine- and deep learning architectures for crash-severity prediction [6], [7], [9]. Ensemble-based classifiers have also been extensively investigated, including Random Forest and other ensemble learning approaches [10], [11], [12], [13], as well as more recent XGBoost, LightGBM, and optimized ensemble models [3], [14], [15], [16], [17]. Class imbalance is another major challenge in traffic accident severity modeling because fatal crashes generally constitute only a small proportion of the available records. Previous studies have addressed this issue using imbalance-aware ensemble learning [3], generative adversarial networks [4], and comparisons between generative and conventional resampling strategies [5]. Data quality and missing information can further affect crash-severity prediction; therefore, imputation and missing-data handling have also received considerable attention in recent research [18], [19], [20]. In parallel, explainable artificial intelligence has increasingly been used to improve the interpretability of accident-risk models, including feature-level risk analysis and SHAP-based explanations [21], [22]. Despite these advances, much of the existing literature focuses primarily on classifier performance, feature selection, imbalance mitigation, or post-hoc model interpretation. Comparatively less attention has been given to interpretable composite risk features whose incremental predictive value is evaluated under strict leakage-aware validation. In the present study, Association Rule Mining (ARM) is used independently to identify interpretable high-risk combinations. In contrast, the Crash Severity Risk Score (CSRS) is constructed separately from training-derived category-risk mappings and evaluated as an additional predictive feature across multiple classifiers. This design enables the explanatory contribution of ARM and the predictive contribution of CSRS to be examined without treating the two components as dependent.

III. MATERIALS AND METHODS

1. Dataset

The dataset used in this study comprises vehicle and accident information obtained from the traffic accident records of the Republic of Türkiye Ministry of Interior, General Directorate of Security. Passenger and pedestrian records were not included at this stage. The integrated dataset covers the 2018–2025 period, while

the Ankara subset covering 2020–2025 was used for the primary modeling experiments. Dataset and partition statistics are summarized in Table 1.

Table 1. Summary information on the dataset and modeling subset

Feature	Value / Description
Data source	Traffic accident records of the General Directorate of Security
Scope	Vehicle and accident information for the 2018–2025 period
Integrated dataset size	2,463,711 rows and 70 variables
Modeling subset	Ankara, 2020-2025
Modeling dataset size	135,776 vehicle-accident records from 79,139 unique accidents
Target variable	0 = Injury accident, 1 = Fatal accident
Row-level target distribution	133,809 injury records (98.551%) and 1,967 fatal records (1.449%)
Accident-level target distribution	77,849 injury accidents (98.37%) and 1,290 fatal accidents (1.63%)
Training Set	108,610 vehicle-accident records from 63,311 unique accidents
Training row-level distribution	107,037 injury records and 1,573 fatal records (1.448% fatal)
Held-out test set	27,166 vehicle-accident records from 15,828 unique accidents
Test row-level distribution	26,772 injury records and 394 fatal records (1.450% fatal)
Train–test accident overlap	0 unique accidents

2. Data Preprocessing

The traffic accident data used in this study consist of integrated vehicle and accident records. Variable names and data types were standardized, and the Ankara subset covering 2020–2025 was extracted for modeling. The binary target variable was defined as 1 for fatal outcomes and 0 for injury outcomes. Variables that directly contain information on the total number of fatalities or injuries were excluded from the explanatory feature set to prevent direct target leakage. Because multiple vehicle records may correspond to the same accident, a conventional row-level random split could assign records from the same accident to both the training and test sets. Moreover, inspection of the data showed that KazaId values could recur across different calendar years. Therefore, a composite accident identifier, AccidentGroupId = (accident year, KazaId), was constructed and used as the grouping unit for data partitioning. The Ankara subset contained 135,776 vehicle–accident records corresponding to 79,139 unique accidents. An 80/20 stratified split was then performed at the unique-accident level using a fixed random seed of 42, with stratification based on the binary accident outcome. All vehicle accident records belonging to the same AccidentGroupId were assigned to the same partition. This resulted in 108,610 records from 63,311 unique accidents in the training set and 27,166 records from 15,828 unique accidents in the held-out test set, with no accident overlap between the two partitions. All data-dependent preprocessing and target-informed procedures were performed without using information from the held-out test set. Logistic Regression, Random Forest, Balanced Random Forest, XGBoost, and LightGBM used one-hot

encoding; missing categorical values were assigned to a dedicated category; unseen categories were ignored during validation and testing; and missing numerical values were imputed using training medians. CatBoost used native categorical inputs and training-median numerical imputation. For CSRS construction, category-level risk mappings for training observations were generated using an out-of-fold procedure. In contrast, the mappings applied to the held-out test set were estimated only from the complete training partition. Similarly, ARM candidate discovery and rule selection were restricted to the training partition. Held-out test outcomes were used exclusively for the final independent evaluation.

3. Method

Association Rule Mining (ARM) and the Crash Severity Risk Score (CSRS) were employed as methodologically separate components: ARM identified multivariable conditions associated with fatal outcomes, whereas CSRS was developed as an interpretable supervised risk feature. ARM was conducted using driver-, vehicle-, road-, environmental-, temporal-, and accident-occurrence-related variables. Because fatal outcomes accounted for only 1.4483% of the training vehicle accident records, candidate discovery was performed on a class-balanced subset that included all 1,573 fatal training records and an equal number of randomly sampled injury records. A minimum-support sensitivity analysis was performed using values of 0.01, 0.02, 0.03, and 0.05, and 0.02 was selected for candidate discovery. Rules with a fatal consequence and confidence ≥ 0.65 were retained. Rules with a fatal consequence and confidence ≥ 0.65 were retained. In the balanced discovery subset, this corresponded to a lift of 1.30 and was used only for candidate screening, not as a population-level fatality probability. Candidate rules were subsequently re-evaluated on the original imbalanced training data, where support, fatal-outcome rate, and lift were recalculated under the natural class distribution. To reduce redundancy, pairwise rules were retained if their training-set lift was at least 30% higher than that of the corresponding stronger single-condition rule. Following training-stage filtering, 58 candidate rules were frozen and evaluated on the held-out test set using lift and one-sided Fisher's exact tests. Secondary stability criteria required preservation of association direction, at least 70% train-to-test lift retention, and support from at least 10 fatal test records. P-values were adjusted using the Benjamini–Hochberg false-discovery-rate procedure. CSRS was constructed independently from ARM using six components: vehicle, driver, road, environment, time, and accident occurrence type. To prevent target leakage, five-fold out-of-fold estimation was performed at the unique-accident level within the training partition. For each fold, category-level risk mappings were estimated using the remaining four folds and then applied to the held-out fold. Category-specific fatal-outcome proportions were compared with the training baseline using 95% Wilson confidence intervals; intervals below, overlapping, or above the baseline were assigned risk scores of 0, 1, and 2, respectively. For components with multiple subvariables, the maximum subvariable score was used to ensure each component remained on the 0–2 scale. The total CSRS for observation *I* was calculated as shown in (1).

$$\mathrm{CSRS_i} = \sum w_j.R(x_{ij}) \tag{1}$$

Here, x_{ij}, denotes the category value of the j-th risk component for observation *i*, $R(x_{ij})$ represents the corresponding risk value (0, 1, or 2), and w_j denotes the component weight. All weights were ss = 1.o w alt j 1.o =*1.*. Consequently, the theoretical CSRS range is 0–12, with higher scores representing risk profiles more strongly associated with fatal outcomes. After out-of-fold CSRS values were generated for the training records, final category-risk mappings were estimated from the complete training partition and applied unchanged to the held-out test set. The incremental contribution of CSRS was evaluated using six classifiers—Logistic Regression, Random Forest, Balanced Random Forest, XGBoost, LightGBM, and CatBoost—each fitted with and without CSRS using identical hyperparameters. Logistic Regression used balanced class weights, the liblinear solver, and 2,000 maximum iterations; Random Forest used 500 trees, balanced class weights, and a minimum leaf size of 2; and Balanced Random Forest used 500 trees with balanced sampling, replacement, and

a minimum leaf size of 2. XGBoost and LightGBM used 500 estimators, a learning rate of 0.05, 0.8 subsampling and feature sampling, and training-derived class weights; maximum depth was 5 for XGBoost, and the number of leaves was 31 for LightGBM. CatBoost used 500 iterations, a learning rate of 0.05, depth 6, Logloss, and automatic class-weight balancing. The random seed was 42. Decision thresholds from 0.01 to 0.99 were selected on an accident-level validation fold from the training partition by maximizing the fatal-class F1-score; ties were resolved by higher recall, then by the lower threshold. Models were then fitted on the complete training partition and evaluated on the held-out test set using the fixed thresholds. After threshold selection, models were fitted using the complete training partition and evaluated on the held-out test set using the previously fixed thresholds. To quantify uncertainty in the CSRS/no-CSRS comparisons, paired bootstrap resampling was performed at the AccidentGroupId level with 1,000 resamples, preserving all vehicle records for each sampled accident. For strict temporal validation, models were trained using records from 2020–2023, decision thresholds were selected using 2024 data, and final performance was evaluated on 2025 records without using 2025 outcomes during model or threshold selection. For geographic validation, Ankara-trained models were evaluated on İstanbul records from 2020 to 2025. İstanbul was selected as the non-Ankara province with the largest number of unique accidents. Ankara-derived decision thresholds and CSRS mappings were applied unchanged. CSRS robustness was assessed using 90%, 95%, and 99% Wilson confidence intervals for category-risk assignment, with 95% used as the primary setting. Class separation and LightGBM discrimination were compared across the three variants using the same leakage-controlled procedure.

4. Performance Metrics

Since fatal outcomes constitute a highly underrepresented minority class, overall accuracy alone was not considered sufficient for model evaluation. Performance was therefore assessed primarily using fatal-class precision, recall (sensitivity), F1 Score, specificity, balanced accuracy, ROC AUC, and Average Precision (AP). False-positive (FP) and false-negative (FN) counts were reported in the main comparison table. ROC-AUC evaluates ranking performance across decision thresholds, whereas Average Precision (AP) summarizes precision–recall performance and was used as a primary threshold-independent metric for the minority fatal class. The main threshold-dependent performance metrics used in this study are given in (2), (3), and (4).

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Decision thresholds were selected using the validation data, as described in Section III.3, and subsequently fixed before evaluation on the corresponding held-out test sets. Consequently, no performance metric or confusion matrix results from the final test data were used for threshold optimization.

IV. EXPERIMENTAL RESULTS

The held-out test set contained 27,166 vehicle–accident records, including 26,772 injury records and 394 fatal records, corresponding to a fatal-outcome prevalence of 1.4503%. No AccidentGroupId was shared between the training and test sets. Figure 1 summarizes the held-out test-set lift values of the strongest pairwise ARM patterns selected exclusively from the training data.

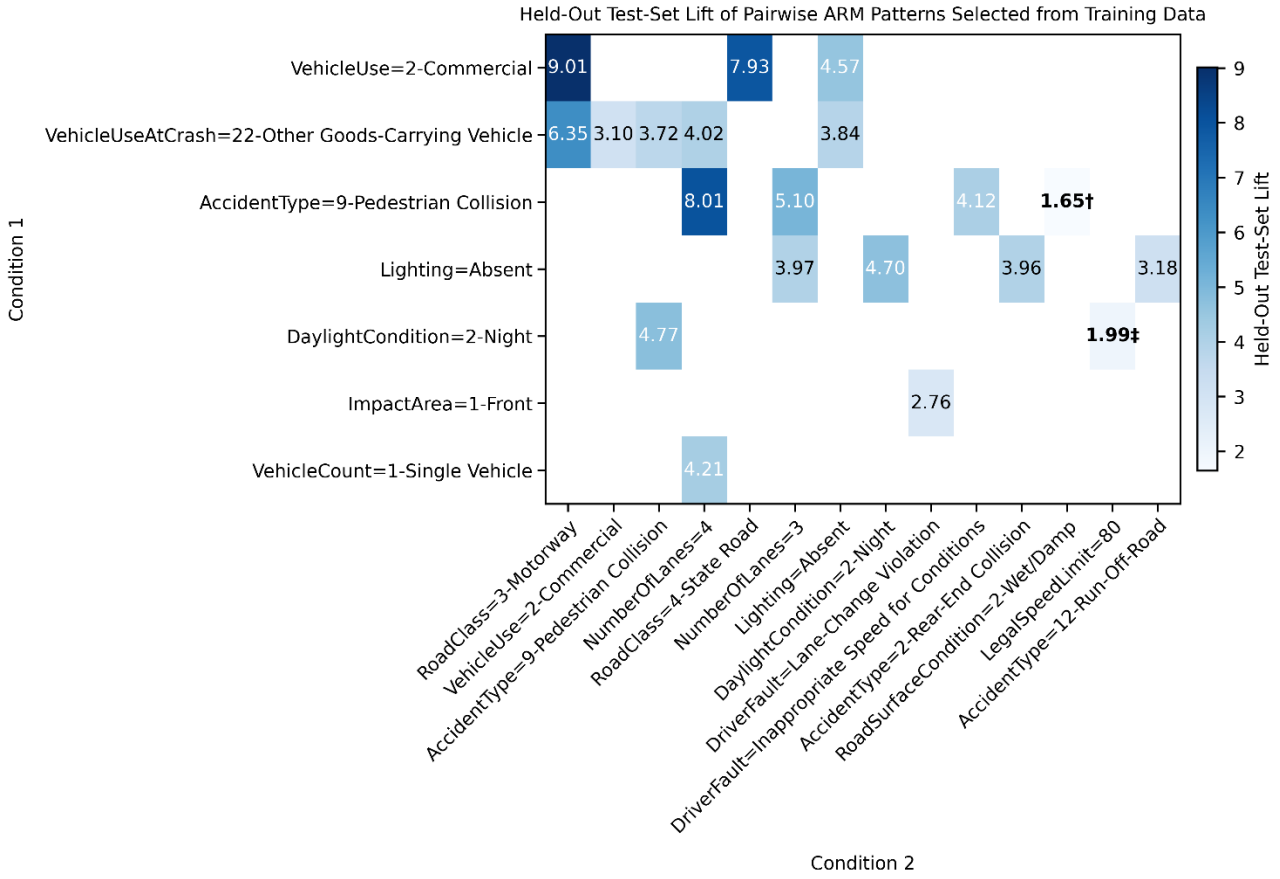


Figure 1. Held-out test-set lift of pairwise ARM patterns selected exclusively from training-set results. Vehicle type was excluded from the visualization only. Nineteen of the 20 displayed patterns remained statistically significant after Benjamini–Hochberg FDR correction applied across all 58 frozen candidate rules, and 18 additionally satisfied the predefined secondary stability criteria. † Not significant after FDR correction and based on fewer than 10 fatal test records. ‡ Significant in the held-out test set but below the 70% train-to-test lift-retention criterion.

The ARM support-threshold sensitivity analysis considered minimum support values of 0.01, 0.02, 0.03, and 0.05, and 0.02 was retained for candidate discovery. The number of fatal-containing candidate itemsets decreased from 2,105 at a minimum support of 0.01 to 566 at 0.05; therefore, 0.02 was retained as a compromise between rule coverage and sparsity, yielding 1,314 candidates. Following training-stage discovery, re-evaluation under the original imbalanced distribution, and redundancy filtering, 58 candidate rules were frozen before examination of the held-out test set. Of these, 52 remained statistically significant after one-sided Fisher's exact testing with Benjamini–Hochberg false-discovery-rate correction, and 51 also satisfied the predefined secondary stability criteria. Among the strongest replicated associations, the combination Vehicle Use = Commercial and Road Class = Motorway occurred in 153 test records, including 20 fatal records, corresponding to a fatal-outcome rate of 13.072%. Relative to the overall test-set fatal prevalence of 1.4503%, this association yielded a test lift of 9.013 (odds ratio = 10.711, 95% CI: 6.621–17.327; FDR-adjusted $q = 3.414 \times 10^{-13}$).

Although one displayed a pattern that did not replicate, and another showed attenuation, most retained elevated associations with fatal outcomes when evaluated independently. Beyond these rule-based associations, CSRS also showed descriptive separation between the two outcomes. In the held-out test set, injury records had a mean CSRS of 5.996 (SD = 1.693; median = 6), whereas fatal records had a higher mean score of 7.718 (SD = 1.595; median = 8). Because test-set CSRS values were generated exclusively from training-derived mappings, held-out outcomes were not used to construct scores. The incremental contribution of CSRS was then evaluated across all six classifiers by comparing otherwise identical models with and without the score. Decision thresholds were selected from validation data and fixed before the held-out test evaluation. The resulting performance measures and FP/FN error counts are presented in Table 2.

Table 2. Held-Out Test Performance Of The Six Classifiers With And Without CSRS

Model	CSRS	Th.	AP	AUC	Prec.	Rec.	F1	FP/FN
LR	No	0.88	0.0952	0.8494	0.1185	0.2614	0.1631	766/291
LR	Yes	0.89	0.0938	0.8471	0.1148	0.2462	0.1566	748/297
RF	No	0.31	0.1100	0.8467	0.1725	0.2132	0.1907	403/310
RF	Yes	0.24	0.1135	0.8440	0.1417	0.3071	0.1939	733/273
BRF	No	0.70	0.1082	0.8511	0.1193	0.3756	0.1810	1093/246
BRF	Yes	0.72	0.1125	0.8522	0.1251	0.3046	0.1774	839/274
XGBoost	No	0.82	0.1317	0.8334	0.1669	0.2640	0.2045	519/290
XGBoost	Yes	0.82	0.1405	0.8411	0.1782	0.2487	0.2076	452/296
LGBM	No	0.75	0.1301	0.8226	0.1611	0.3046	0.2107	625/274
LGBM	Yes	0.78	0.1462	0.8302	0.1802	0.2817	0.2198	505/283
CatBoost	No	0.71	0.1243	0.8247	0.1108	0.4036	0.1739	1276/235
CatBoost	Yes	0.76	0.1258	0.8310	0.1346	0.3325	0.1917	842/263

Th.: validation-selected decision threshold; AP: average precision; AUC: area under the receiver operating characteristic curve (ROC-AUC); Prec.: fatal-class precision; Rec.: fatal-class recall; F1: fatal-class F1-score; FP: false-positive count; FN: false-negative count; LR: logistic regression; RF: random forest; BRF: balanced random forest; LGBM: LightGBM.

As shown in Table 2, CSRS improved Average Precision in five of the six classifiers and fatal-class F1-score in four. LightGBM with CSRS achieved the highest AP (0.1462) and fatal-class F1-score (0.2198), while Balanced Random Forest with CSRS achieved the highest ROC-AUC (0.8522). CatBoost without CSRS produced the highest fatal-class recall (0.4036). Table 2 also reports false-positive and false-negative counts; since the held-out test set contained 26,772 injury and 394 fatal records, the corresponding true-negative and true-positive counts can be derived directly from these values. Paired accident-level bootstrap analysis with 1,000 resamples showed that no CSRS-related F1-score difference had a 95% confidence interval excluding zero. In contrast, the clearest threshold-independent improvement was observed for LightGBM AP (+0.0161, 95% CI: 0.0019–0.0326). Generalizability was further examined using strict temporal and geographic validation. In the temporal analysis, models were developed using 2020–2023 data, thresholds were selected in 2024, and 2025 was reserved for final testing. CSRS increased Average Precision in all six classifiers, with the largest gains observed for XGBoost (0.0978 to 0.1138) and LightGBM (0.1040 to 0.1151); LightGBM with CSRS achieved the highest temporal AP (0.1151) and fatal-class F1-score (0.1876). Geographic validation was conducted by training in

Ankara and evaluating in Istanbul. Under this distribution shift, the highest AP was obtained by XGBoost without CSRS (0.0555), whereas Balanced Random Forest achieved the highest ROC-AUC with CSRS (0.7760). A sensitivity analysis using 90%, 95%, and 99% Wilson confidence intervals showed that the CSRS class-separation pattern remained stable across alternative scoring boundaries. For LightGBM, Average Precision ranged from 0.1387 to 0.1490 and ROC-AUC from 0.8238 to 0.8302.

V. DISCUSSION

The results indicate that ARM and CSRS provide complementary forms of information for traffic accident severity analysis. ARM identified interpretable combinations of conditions associated with elevated fatal-outcome risk, and most of the strongest training-selected patterns retained their direction and elevated lift under held-out evaluation. CSRS, which was constructed independently of ARM, showed a clear descriptive separation between injury and fatal records and improved Average Precision across five of the six evaluated classifiers. However, its predictive contribution was model-dependent rather than uniformly beneficial across all metrics. LightGBM obtained the strongest overall combination of AP and fatal-class F1-score when CSRS was included. In contrast, CatBoost without CSRS achieved the highest fatal-class recall, illustrating the trade-off between detecting fatal cases and limiting false-positive predictions. The bootstrap analysis further suggests that the incremental benefit of CSRS should be interpreted with caution: the clearest statistically supported improvement was observed for LightGBM AP. At the same time, the F1-score differences did not have confidence intervals that excluded zero. Performance also decreased with shifts in temporal and geographic distributions, highlighting the challenge of transferring crash-severity models across time and regions. Nevertheless, the CSRS sensitivity analysis produced similar discrimination results under alternative scoring boundaries, supporting the robustness of the overall scoring strategy. These findings suggest that CSRS is best viewed as an interpretable complementary feature rather than a universally performance-enhancing predictor.

VI. CONCLUSION AND FUTURE WORK

This study presented an interpretable and leakage-aware machine learning framework for traffic accident severity prediction by combining ARM-based risk-pattern discovery with an independently constructed Crash Severity Risk Score (CSRS). The held-out results showed that fatal records had higher CSRS values than injury records, while the predictive contribution of CSRS varied across classifiers. LightGBM with CSRS achieved the highest held-out Average Precision and fatal-class F1-score, and bootstrap analysis supported a modest but measurable improvement in its ranking performance. Temporal and cross-province validation further showed that model performance decreased under distribution shift, although CSRS retained useful predictive information in several models. Future work will focus on explaining model decisions using XAI methods such as SHAP and extending the framework toward spatial road-segment risk modeling, graph-based learning, and safe-route recommendation.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Republic of Türkiye Ministry of Interior, General Directorate of Security, for providing the traffic accident records used in this study. The analyses and interpretations presented in this paper are solely the responsibility of the authors.

REFERENCES

[1]"Road traffic injuries." Accessed: Nov. 30, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[2]P. B. Silva, M. Andrade, and S. Ferreira, "Machine learning applied to road safety modeling: A systematic literature review," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 7, no. 6, pp. 775–790, Dec. 2020, doi: 10.1016/J.JTTE.2020.07.004.

[3]K. Aziz, F. Chen, M. Ahmad, M. S. Khan, M. M. Sabri Sabri, and H. Almujibah, "An interpretable dynamic ensemble selection multiclass imbalance approach with ensemble imbalance learning for predicting road crash injury severity," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-08935-x.

[4]J. Chen, Z. Pu, N. Zheng, X. Wen, H. Ding, and X. Guo, "A novel generative adversarial network for improving crash severity modeling with imbalanced data," *Transp. Res. Part C Emerg. Technol.*, vol. 164, p. 104642, Jul. 2024, doi: 10.1016/J.TRC.2024.104642.

[5]B. Zhou, Q. Zhou, and Z. Li, "Addressing Data Imbalance in Crash Data: Evaluating Generative Adversarial Networks' Efficacy Against Conventional Methods," *IEEE Access*, vol. 13, pp. 2929–2944, 2025, doi: 10.1109/ACCESS.2024.3524620.

[6]S. Zhang, A. Khattak, C. M. Matara, A. Hussain, and A. Farooq, "Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents," *PLoS One*, vol. 17, no. 2, p. e0262941, Feb. 2022, doi: 10.1371/JOURNAL.PONE.0262941.

[7]M. Manzoor *et al.*, "RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model," *IEEE Access*, vol. 9, pp. 128359–128371, 2021, doi: 10.1109/ACCESS.2021.3112546.

[8]K. Santos, J. P. Dias, and C. Amado, "A literature review of machine learning algorithms for crash injury severity prediction," *J. Safety Res.*, vol. 80, pp. 254–269, 2022, doi: 10.1016/j.jsr.2021.12.007.

[9]M. Zheng *et al.*, "Traffic accident's severity prediction: A deep-learning approach-based CNN network," *IEEE Access*, vol. 7, pp. 39897–39910, 2019, doi: 10.1109/ACCESS.2019.2903319.

[10]S. Ceven and A. Albayrak, "Traffic accident severity prediction with ensemble learning methods," *Computers & Electrical Engineering*, vol. 114, 2024, doi: 10.1016/j.compeleceng.2024.109101.

[11]T. Whasphutthisit and W. Jitsakul, "Ensemble Learning Approach for Enhanced Road Deaths Prediction," in *2022 International Conference on Digital Government Technology and Innovation (DGTi-CON)*, 2022, pp. 15–19. doi: 10.1109/DGTi-CON53875.2022.9849202.

[12]R. E. Almamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology, JEEIT 2019 - Proceedings*, pp. 272–276, May 2019, doi: 10.1109/JEEIT.2019.8717393.

[13]M. Yan and Y. Shen, "Traffic Accident Severity Prediction Based on Random Forest," *Sustainability 2022, Vol. 14, Page 1729*, vol. 14, no. 3, p. 1729, Feb. 2022, doi: 10.3390/SU14031729.

[14]M. S. Alam *et al.*, "Traffic Crash-Severity Prediction using Interpretable Optimized Ensemble Models," *International Journal of Intelligent Transportation Systems Research*, vol. 23, no. 3, pp. 1585–1607, 2025, doi: 10.1007/s13177-025-00531-7.

[15]F. N. Nur, M. N. Islam, and A. S. Afridi, "AI-Enabled Ensemble Learning Framework for Predicting Road Accident Severity," in *2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025*, 2025. doi: 10.1109/QPAIN66474.2025.11172248.

[16]F. Chen *et al.*, "Traffic accident severity prediction based on an enhanced MSCPO-XGBoost hybrid model," *Scientific Reports,* vol. 15, no. 25729, 2025, doi: 10.1038/s41598-025-00797-7.

[17]W. Tang, D. Wang, X. Yang, and G. Li, "Predicting and Analyzing Accident Severity with Improved LightGBM Method," in *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 2023, pp. 2031–2034. doi: 10.1109/ICSP58490.2023.10248671.

[18] X. Gao *et al.*, "Reliable imputation of incomplete crash data for predicting driver injury severity," *Accid. Anal. Prev.*, vol. 216, p. 108020, Jun. 2025, doi: 10.1016/J.AAP.2025.108020.

[19] Y. Zhang, X. Kong, W. Zhou, J. Liu, Y. Fu, and G. Shen, "A Comprehensive Survey on Traffic Missing Data Imputation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 19252–19275, 2024, doi: 10.1109/TITS.2024.3478816.

[20] L. Li, C. G. Prato, and Y. Wang, "Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and random forest classifier," *Accid. Anal. Prev.*, vol. 146, p. 105744, Oct. 2020, doi: 10.1016/J.AAP.2020.105744.

[21] I. Abdulrashid, R. Zanjirani Farahani, S. Mammadov, M. Khalafalla, and W. C. Chiang, "Explainable artificial intelligence in transport Logistics: Risk analysis for road accidents," *Transp. Res. E Logist. Transp. Rev.*, vol. 186, p. 103563, Jun. 2024, doi: 10.1016/J.TRE.2024.103563.

[22] O. I. Aboulola, "Improving traffic accident severity prediction using MobileNet transfer learning model and SHAP XAI technique," *PLoS One*, vol. 19, no. 4 April, 2024, doi: 10.1371/journal.pone.0300640.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Cloud-CTI-Engine: An Integrated AWS Security Posture and Threat Intelligence Platform

Gerta Hodolli
Department of Computer Engineering
University of Prishtina
Prishtina, Kosovo
gerta.hodolli@student.uni-pr.edu

Rrezearta Thaqi
Department of Computer Engineering
University of Prishtina
Prishtina, Kosovo
rrezearta.thaqi@uni-pr.edu

Blerim Rexha
Department of Computer Engineering
University of Prishtina
Prishtina, Kosovo
blerim.rexha@uni-pr.edu

Abstract

AWS security evidence is fragmented across configuration state, identity policies, network exposure rules, audit logs, threat intelligence feeds, and compliance controls, which makes it difficult for analysts to connect cloud misconfigurations with runtime activity and external threat context. This paper presents Cloud-CTI-Engine, a locally executable framework that brings AWS posture findings, CloudTrail events, threat intelligence enrichment, and compliance mapping into a single assessment workflow. The system implements deterministic checks for S3, IAM, and EC2; extracts indicators of compromise from CloudTrail records; enriches IP addresses, domains, and file hashes via external intelligence services; and maps findings to the CIS AWS Foundations Benchmark, NIST Cybersecurity Framework, ISO/IEC 27001, and SOC 2. Unlike isolated posture scanners or log-analysis tools, it normalizes these evidence sources into a common schema for findings. It executes entirely on the analyst's machine, without a centralized service or external credential storage. The platform was evaluated on controlled AWS accounts with contrasting security postures. It produced posture scores of 80/100 and 52/100 across 30 mapped controls, with 7 and 16 control failures, respectively, and extracted 8 indicators from 45 CloudTrail events, of which 6 were enriched, and 2 were flagged as suspicious or malicious.

Keywords: cloud security posture management, AWS security, CloudTrail analysis, threat intelligence, indicators of compromise, compliance mapping, cloud misconfiguration

I. INTRODUCTION

Cloud computing has transformed how organizations deploy and manage IT infrastructure. Platforms such as Amazon Web Services (AWS) provide scalable and flexible environments but introduce new cybersecurity challenges due to complex configurations and distributed architectures [1]. Misconfigurations such as publicly exposed storage buckets, excessive IAM privileges, or insecure network configurations increase the attack surface. At the same time, threat intelligence provides information about malicious infrastructure, attack patterns, and indicators of compromise [2]. AWS configuration data, IAM policy state, security group exposure, CloudTrail activity, external IOC reputation, and compliance controls are usually analyzed through separate workflows. In addition, organizations must ensure their cloud environments comply with established security frameworks, including the CIS AWS Foundations Benchmark, NIST Cybersecurity Framework, ISO/IEC 27001, and SOC 2. Mapping detected security findings to these controls is essential for evaluating compliance posture and prioritizing remediation efforts [3]. This paper presents Cloud-CTI-Engine, a local-first AWS security assessment framework that integrates cloud posture checks, CloudTrail analysis, threat intelligence enrichment, and compliance mapping into a unified workflow. The system normalizes posture findings, log-derived indicators, enrichment results, severity levels, and framework mappings into a common finding

structure. This allows analysts to move from isolated observations, such as a public S3 bucket, an exposed EC2 security group, or a suspicious CloudTrail source IP, toward a consolidated assessment that connects technical evidence with operational and compliance context. The main contributions of this paper are as follows:

- A local-first AWS security assessment architecture that combines deterministic posture checks, CloudTrail parsing, threat intelligence enrichment, and compliance mapping in a single workflow.
- A normalized finding model for representing AWS misconfigurations, log-derived indicators, enrichment metadata, severity, supporting evidence, and framework mappings.
- A prototype implementation covering selected AWS services, including S3, IAM, and EC2, together with CloudTrail IOC extraction and enrichment through external intelligence sources.
- A scenario-based evaluation on controlled AWS accounts with contrasting security postures, reporting posture scoring, severity distribution, compliance mapping outcomes, and indicator enrichment results.

II. RELATED WORK

Cloud security posture management (CSPM) has emerged as a key approach for continuously assessing cloud configurations and identifying misconfigurations, compliance violations, and security risks [4]. Prior work shows that CSPM can reduce exposure caused by public storage, excessive IAM permissions, and insecure network rules. However, most CSPM approaches focus primarily on configuration state, whereas separate tools typically handle runtime activity and threat intelligence. AWS-native services also provide partial coverage. AWS Config supports configuration tracking and compliance rules, AWS Security Hub aggregates security findings, and Amazon GuardDuty analyzes activity logs to detect suspicious behavior. Open-source tools such as Prowler and ScoutSuite provide posture and compliance checks, while Cloudsplaining focuses on IAM policy risks. These tools are useful, but they do not provide a local-first workflow that directly combines posture findings, CloudTrail-derived indicators, threat intelligence enrichment, and compliance mapping into one normalized structure. Threat intelligence and CloudTrail analysis have also been studied separately. Threat intelligence adds context about indicators such as IP addresses, domains, URLs, and file hashes, while CloudTrail analysis helps identify suspicious AWS activity. Cloud-CTI-Engine connects these areas by using CloudTrail parsing to extract indicators and correlate them with posture findings, enrichment results, and compliance controls. Table 1 compares the proposed platform with representative AWS-native and open-source approaches.

Table 1. Comparison of related AWS security assessment approaches, based on published product and project documentation as of July 2026

Approach	Posture checks	CloudTrail analysis	Threat intelligence enrichment	Compliance mapping	Unified local workflow
AWS Config	Yes	Limited	No	Yes	No
AWS Security Hub	Yes	Partial	Limited	Yes	No
Amazon GuardDuty	No	Yes	Yes	No	No
Prowler / ScoutSuite	Yes	Limited or no	No	Yes	Partial
Cloudsplaining	IAM-focused	No	No	Limited	Partial

Cloud-CTI-Engine	Yes	Yes	Yes	Yes	Yes

III. BACKGROUND

Cloud computing delivers computing resources as on-demand services over the internet, removing the need to manage physical infrastructure [5]. Regulatory compliance is more challenging in cloud environments, particularly when cross-border data transfers and differing international regulations are involved [6]. The shared responsibility model divides security duties between the cloud service provider and the customer [7]. Effective monitoring requires native logging such as AWS CloudTrail, centralized log correlation, and anomaly detection [8]. Threat intelligence provides continuous streams of indicators of compromise, such as malicious domains and IP addresses [9]. Mapping detected threats to frameworks such as the NIST Cybersecurity Framework and CIS Controls supports compliance verification [10]. AI-based approaches on cloud log sources detect behavioral anomalies with higher accuracy than signature-based methods [11]. Combining threat intelligence with machine learning is effective for volumetric attacks such as DDoS [12].

IV. SYSTEM ARCHITECTURE

Cloud-CTI-Engine follows a modular architecture composed of three main components: a command-line interface (CLI), a backend API server, and a web-based frontend. The CLI performs the main security analysis tasks, the backend orchestrates communication between the CLI and the graphical interface, and the frontend provides visualization and user interaction. The Python-based CLI implements the core analysis engine. It performs cloud security posture checks for selected AWS services, including S3, IAM, and EC2, processes CloudTrail logs, and enriches indicators such as IP addresses, domains, and file hashes. It interacts with AWS services through the boto3 SDK, uses Typer for command orchestration and parameter handling, and applies Pydantic to validate structured data and normalize results into a consistent schema. The backend API server, implemented in Node.js with Express, serves as a lightweight orchestration layer. It does not perform security analysis directly; instead, it invokes the CLI as a subprocess with the required parameters and returns the results to the frontend. This design centralizes the analysis logic in the CLI while allowing the API to manage configuration, execution workflows, and result retrieval. The web frontend allows users to initiate scans, visualize analysis results, and review detected findings. The platform supports direct CLI execution and web platform mode. In platform mode, the frontend sends HTTP requests to the API, which launches the CLI, performs the analysis, and returns the results. All processing is performed locally on the user's machine. The platform does not require a centralized backend service or external credential storage, ensuring that AWS credentials and analysis data remain under the user's control.

1. Cloud Posture Check

Cloud posture assessment uses deterministic security controls for Amazon S3, Identity and Access Management (IAM), and Elastic Compute Cloud (EC2), representing data storage, identity management, and network security. Each modular control receives an authenticated AWS session, queries the relevant configuration via AWS APIs, and evaluates it against predefined rules informed by recommended security practices and industry frameworks. Detected issues are normalized into a unified finding structure containing the service name, resource identifier, issue type, severity, and supporting evidence. An AWS scan runner initializes the session, resolves the target services and regions, executes the service-specific checks, and aggregates the findings into a consolidated posture report.

- S3 Security Checks

The S3 module performs four configuration checks for each bucket: public access detection, default encryption validation, versioning status verification, and server access logging configuration.

Public access detection evaluates three configuration components: Block Public Access (BPA) settings, bucket policies, and access control lists (ACL). A high-severity finding is generated when BPA protections are not fully enabled, and either the bucket policy or ACL configuration allows public access. The default encryption check verifies that buckets use approved encryption mechanisms, such as AES-256 or AWS Key Management Service (AWS KMS). Buckets that do not have encryption enabled or use unsupported encryption algorithms generate a medium-severity finding. Versioning and server access logging checks are configurable controls that can be enabled or disabled through the platform configuration. When enabled, buckets without versioning or logging configured generate low-severity findings.

- IAM Security Checks

The IAM module evaluates identity and access configurations within the AWS account. It performs five security checks: multi-factor authentication (MFA) for IAM users, root account MFA configuration, access key rotation policies, detection of unused access keys, and policy analysis for administrative or risky permissions. For each IAM user, the platform verifies whether an MFA device is configured. Users without MFA are reported as high-severity findings due to the increased risk of credential compromise. Access keys are evaluated against configurable thresholds that define the maximum allowed key age and maximum unused period. Keys that exceed the configured age limit are flagged for rotation, while keys that have never been used or have not been used within the defined threshold are classified as unused. Root account MFA is evaluated once at the account level. If MFA is not enabled for the root account, the system generates a finding indicating a critical security risk. Policy analysis consists of two separate checks. The first identifies administrative wildcard permissions, which occur when policies grant unrestricted access using Action: "*" and Resource: "*". The second check identifies risky policy constructs, such as NotAction and NotResource used in Allow statements, as well as configurable privilege-escalation actions, such as iam: CreateAccessKey and iam: AttachUserPolicy.

- EC2 Security Checks

The EC2 module analyzes security groups across AWS regions using two checks: public exposure and unused security groups. Inbound rules are evaluated to determine whether sensitive ports are accessible from unrestricted ranges such as 0.0.0.0/0 or::/0. Exposure of all ports is classified as critical, while exposure of specific sensitive ports generates high-severity findings. The module also identifies security groups that are not attached to any network interface. These resources may represent residual or improperly cleaned-up infrastructure. Security groups are enumerated across all available regions, and their inbound rules are inspected for disallowed CIDR ranges and exposed ports.

2. CloudTrail Analysis

CloudTrail logs are analyzed to identify suspicious activity and extract indicators of compromise (IOCs). The analysis engine parses CloudTrail event records and extracts relevant attributes, including sourceIPAddress, eventName, eventSource, and userIdentity. These attributes provide visibility into who performed an action,

which service was accessed, and from which network location the request originated. The platform supports two complementary analysis modes: deterministic baseline analysis and optional AI-assisted log interpretation.

- Baseline IOC Extraction

Baseline analysis performs deterministic extraction of indicators using structured parsing and pattern-matching techniques. The system processes each CloudTrail event and extracts attributes that may indicate security relevance.

- AI-Assisted Log Analysis

In addition to deterministic parsing, the platform optionally performs AI-assisted analysis of CloudTrail activity. CloudTrail events are normalized and passed to a language model that generates structured summaries describing key actors, suspicious actions, and chronological activity patterns.

3. Threat Intelligence

The threat intelligence module evaluates the reputation of indicators extracted from CloudTrail logs or provided manually by the user. The system supports three indicator types: IP addresses, domain names, and file hashes. IP addresses are analyzed using external intelligence services such as AbuseIPDB and IPInfo, which provide metadata including abuse confidence scores, autonomous system numbers (ASN), and hosting provider information. Domain lookups return organizational and geographic metadata, while file hash analysis uses VirusTotal to retrieve malware detection statistics and threat classification results. IP lookups use the AbuseIPDB API v2, while domain lookups resolve the hostname to an IPv4 address and query IPInfo for organizational and geographic metadata. File hash analysis relies on the VirusTotal API v3 and supports common hash algorithms, including MD5, SHA-1, and SHA-256. An optional cloud attribution mechanism re-enriches IP indicators and domains by resolving their IP addresses using IPInfo data. The system compares ASN, organization, and hostname attributes against configurable cloud provider profiles (e.g., AWS, Azure, and GCP) to determine whether an indicator is likely associated with cloud infrastructure. All intelligence queries use a configurable timeout and retry policy with exponential backoff for server-side errors (5xx). The results are normalized into the same ScanResult structure used by the cloud posture checks, ensuring consistent correlation, visualization, and reporting across the platform.

V. COMPLIANCE MAPPING

Compliance evaluation is derived from the findings produced by the cloud posture checks. Each finding is mapped to security controls defined by established frameworks, including the CIS AWS Foundations Benchmark, NIST Cybersecurity Framework, ISO/IEC 27001, and SOC 2. The platform maintains a mapping between finding identifiers and framework control identifiers. When a finding corresponds to a control rule, the associated control is marked as failed. Compliance results are aggregated per framework (passed count, failed count, and percentage of controls passed), providing a high-level overview of how well the evaluated cloud environment aligns with each framework. The UI allows you to expand a framework to see which controls passed or failed and which findings contributed to failures.

VI. EVALUATION



Cloud-CTI-Engine was evaluated using multiple controlled AWS test cases with different security posture conditions. The goal was to verify whether the platform could distinguish between better-configured and more misconfigured environments, generate posture scores, summarize findings by severity, and map results to compliance frameworks. The test cases included AWS accounts with relatively secure configurations and accounts containing intentionally introduced misconfigurations across S3, IAM, and EC2. Table 2 summarizes some of the evaluation results.

Table 2. Scenario-based evaluation results of Cloud-CTI-Engine

Test case	**Evaluation focus**	**Main input**	**Main result**	**Compliance result**
Case 1	Better-configured AWS account	S3, IAM, EC2 scan	80 / 100 posture score; 0 critical, 1 high, 1 medium, 0 low findings	23 passed / 7 failed
Case 2	Misconfigured AWS account	S3, IAM, EC2 scan	52 / 100 posture score; 2 critical, 5 high, 4 medium, 2 low findings	14 passed / 16 failed
Case 3	CloudTrail and threat intelligence analysis	45 CloudTrail events and 8 indicators	8 indicators extracted; 6 enriched; 2 marked suspicious or malicious	Not applicable
Case 4	Extended misconfiguration (10 known weaknesses)	S3, IAM, EC2 scan	44/100; 3 critical, 7 high, 5 medium, 3 low	11/19

Each test case was prepared with a documented baseline before scanning. Cases 2 and 4 included deliberately introduced misconfigurations: Case 2 contained seven weaknesses (a public S3 bucket, missing encryption, two users without MFA, a stale access key, a wildcard IAM policy, and two exposed EC2 security groups); Case 4 extended this with three further weaknesses (a second public bucket, an unused access key, and an RDP-exposed security group), totalling ten. Case 3 used a fixed set of 45 CloudTrail events, each containing eight pre-identified indicator values (IP addresses, domains, and file hashes), two of which were sourced from known malicious entries in the enrichment services. For Cases 2 and 4, all 13 introduced misconfigurations were detected (zero false negatives). Case 1 produced two additional findings (one high, one medium) not part of any introduced weakness; manual review confirmed both as genuine residual issues in the account baseline, not false positives. For Case 3, all eight indicators were correctly extracted, and both pre-identified malicious entries were flagged as suspicious or malicious by the enrichment services, with no benign indicators misclassified. Posture scores in Cases 2 and 4 are lower than the count of introduced weaknesses alone would suggest, because the scoring algorithm also accounts for passing controls, causing the score to drop nonlinearly.

Figure 1 shows the summary dashboard generated during the evaluation for Case 1. It presents the main posture and compliance results, while the platform also provides detailed views for individual findings, compliance controls, CloudTrail analysis, AI-assisted insights, and threat intelligence enrichment.

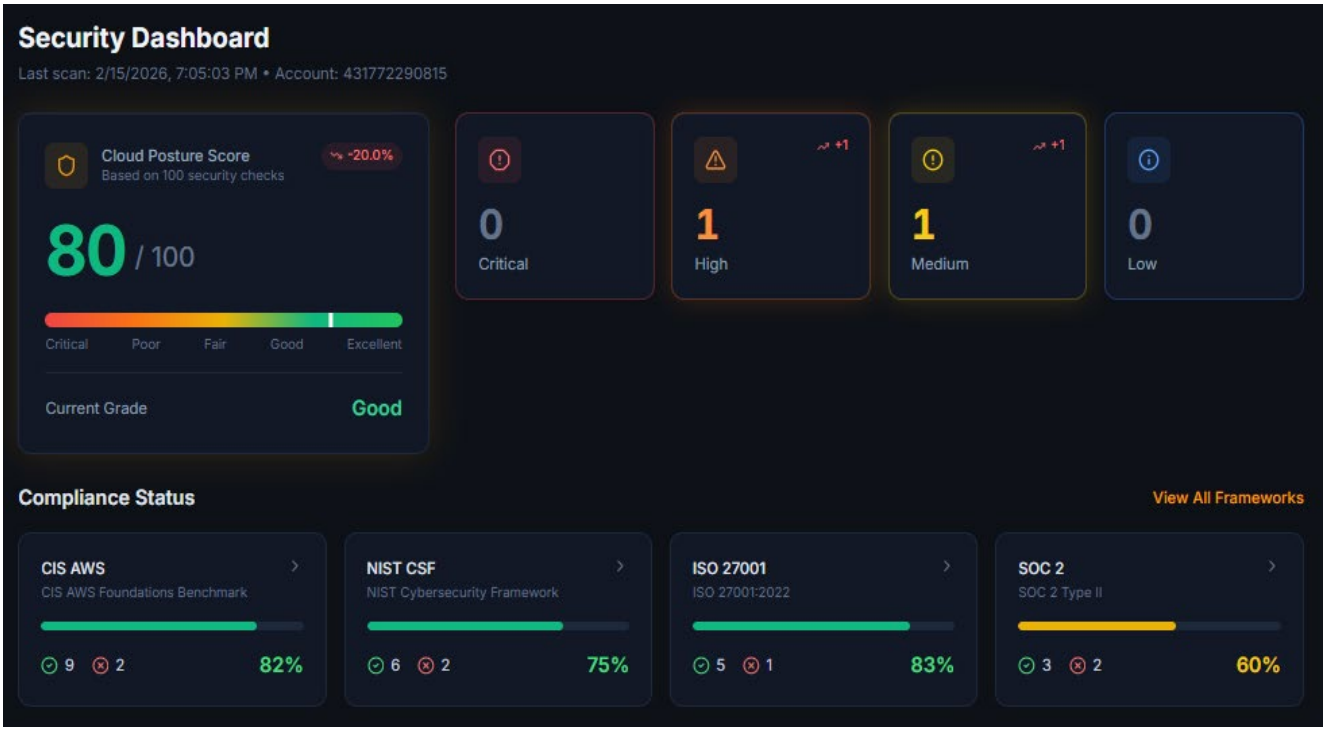


Figure 1. Summary dashboard of Cloud-CTI-Engine showing posture scoring, severity distribution, and compliance mapping results.

The results show that Cloud-CTI-Engine produces different posture scores and distribution of findings based on the security state of the evaluated AWS account. The better-configured account achieved a score of 80 out of 100, with no critical findings and few high and medium findings, while the more misconfigured account produced more findings. The CloudTrail and threat intelligence test also showed that the platform can extract indicators from activity data and enrich them using external intelligence sources.

VII. LIMITATIONS

The evaluation presented in this paper is scenario-based and demonstrative rather than statistical. It uses a small number of controlled AWS accounts, and the number of misconfigurations deliberately introduced in each account is not established as ground truth. Detection coverage, false-positive rates, and runtime overhead are therefore not quantified, and the reported posture scores should be read as indicative of relative configuration quality rather than as validated measurements. Posture coverage is also limited to selected controls for Amazon S3, IAM, and EC2 within a single cloud provider. Hence, the compliance mapping reflects only those controls that the implemented checks can evaluate. Threat intelligence enrichment depends on external services, including AbuseIPDB, IPInfo, and VirusTotal. Each service enforces rate limits: AbuseIPDB limits requests to 1,000 per day on the free tier, IPInfo provides 50,000 requests per month, and VirusTotal enforces 4 requests per minute and 500 per day on the free public API. The platform applies configurable per-request timeouts and exponential backoff for 5xx errors, but does not implement token-bucket rate limiting or graceful degradation when a service is unavailable.

The comparison presented in Table 1 is derived from published product and project documentation. The tools were not executed against the same test accounts, so the table reflects advertised functional coverage rather than measured performance.

The AI-assisted log interpretation mode is an optional, preliminary feature that is not evaluated in this work and should not be relied upon for production security analysis. Assessing the faithfulness of generated summaries, the risk of hallucinated content in a security analysis context, and the impact of prompt design on output quality are left to future work. Deterministic baseline extraction remains the primary and only evaluated indicator source and does not depend on this component.

VIII. CONCLUSION

This paper presented Cloud-CTI-Engine, a locally executable platform that integrates cloud posture analysis, CloudTrail inspection, threat intelligence enrichment, and compliance mapping. The evaluation showed that it can detect AWS misconfigurations, extract and enrich indicators, and map findings to security frameworks. Its CLI-first design supports local security workflows, while the web interface improves accessibility and visualization.

IX. FUTURE WORK

Future work will extend the platform to Microsoft Azure and Google Cloud Platform, expand posture checks, and improve AI-assisted analysis automation. Further improvements will include refined cloud activity detection, stronger indicator correlation across data sources, and enhanced reporting for operational and compliance use cases.

ACKNOWLEDGEMENT

The authors express their sincere appreciation to the Kosovo Cyber Team for supporting this research.

REFERENCES

[1] A. B. Nassif, M. A. Talib, Q. Nasir, H. Albadani, and F. M. Dakalbab, “Machine learning for cloud security: A systematic review,” IEEE Access, vol. 9, pp. 20717–20735, 2021. doi: 10.1109/ACCESS.2021.3054129.

[2] W. Tounsi and H. Rais, “A survey on technical threat intelligence in the age of sophisticated cyber attacks,” Computers & Security, vol. 72, pp. 212–233, Jan. 2018. doi: 10.1016/j.cose.2017.09.001.

[3] B. Reuben-Owoh and E. Haig, “A systematic review of voluntary cybersecurity standards and frameworks,” International Journal of Information Security, vol. 24, Art. no. 206, Sep. 2025. doi: 10.1007/s10207-025-01121-0.

[4] F. N. U. Jimmy, “Cloud security posture management: Tools and techniques,” Journal of Knowledge Learning and Science Technology, vol. 2, no. 3, pp. 619–636, 2023. doi: 10.60087/jklst.vol2.n3.p622.

[5] Google, “What is cloud computing?” Google Cloud. [Online]. Available: https://cloud.google.com/learn/what-is-cloud-computing

[6] Orca Security, “What is cloud security?” Orca Security Blog. [Online]. Available: https://orca.security/resources/blog/what-is-cloud-security/

[7] Palo Alto Networks, “Cloud security is a shared responsibility,” Palo Alto Networks Cyberpedia. [Online]. Available: https://www.paloaltonetworks.com/cyberpedia/cloud-security-is-a-shared-responsibility

[8] O. Alade, "Cloud security in practice: A technical guide to confidentiality, integrity, and availability," World Journal of Advanced Research and Reviews (WJARR), vol. 26, no. 2, pp. 2165–2171, May 2025. doi: 10.30574/wjarr.2025.26.2.1904.

[9] EC-Council, "What is cyber threat intelligence?" EC-Council Cybersecurity Exchange. [Online]. Available: https://www.eccouncil.org/cybersecurity-exchange/threat-intelligence/what-is-cyber-threat-intelligence/

[10] J. Toro-Jarrin, P. Eridel, and S. A. Deeb, "Knowledge graph to facilitate cloud security compliance with NIST cybersecurity framework," in Proc. 2024 2nd International Conference on Computer, Vision and Intelligent Technology (ICCVIT), 2024, pp. 1–6. doi: 10.1109/ICCVIT63928.2024.10872451.

[11] P. R. K. Bollipalli, "Enhancing cloud security in AWS using AI-powered anomaly detection and predictive analytics," International Research Journal of Innovations in Engineering and Technology (IRJIET), vol. 9, no. 3, pp. 213–220, Mar. 2025. doi: 10.47001/irjiet/2025.903028.

[12] R. Thaqi, G. Lenzini, M. Völp, and B. Rexha, "Online machine learning for proactive threat intelligence and timely DDoS attack prediction," Journal of Cloud Computing, Springer, Feb. 2026. doi: 10.1186/s13677-026-00848-y.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Context-Aware Regular Expressions for Sensitive Data Detection

Enes Can Güven
Security Development Hub
Trendyol Group
Istanbul, Türkiye
enescan.guven@trendyol.com

Burak Tahtacı
Security Development Hub
Trendyol Group
Istanbul, Türkiye
burak.tahtaci@trendyol.com

Tuna Çakar
Computer Engineering Dept.
MEF University
Istanbul, Türkiye
cakart@mef.edu.tr

Abstract
Regular expressions reliably match fixed formats, but they produce many false positives when used to find sensitive data in noisy corpora such as source code, logs, tickets, and configuration files. We make context a mandatory part of the matching condition: a candidate value is reported only when a keyword, field label, or linguistic cue for its data type, such as card, api key, authorization, or IBAN, occurs within a bounded window of the value. We describe the rule model, pattern families, and implementation choices, and then evaluate the method through paired ablation on two public benchmarks. On CredData, requiring a context anchor raised the precision of generic credential rules from 0.003 to 0.475 and reduced the reviewed volume from 5.6 million lines to 7,235, while recall fell from 0.956 to 0.356. On ai4privacy, precision more than doubled for categories whose value formats collide and was unchanged for self-distinctive formats such as IBANs. Anchoring, therefore, buys precision at a measured cost to recall, so anchored rules are best deployed as a high-confidence alerting layer alongside a high-recall sweep rather than as a replacement for one.

Keywords: regular expressions, context-aware matching, secret scanner, PII scanner

I. INTRODUCTION

Detecting sensitive and personally identifiable data is a recurring requirement in application security, privacy engineering, and compliance tooling. Organizations routinely scan code repositories, CI/CD artifacts, logs, and documents for exposed secrets, personal data, and financial identifiers, and regular expressions remain the most widely deployed component of these scanners. Yet most deployed patterns match on format alone, which causes a persistent precision problem [1], [2]: a format-only expression can capture the lexical structure of a credit card number or an API key perfectly and still not tell whether the matched text denotes that entity or a benign identifier, a synthetic example, a hash, or test data. For operational teams, the consequence is noisy findings, wasted triage effort, and reduced trust in the scanner [3]. For many sensitive-data categories, the value format alone is not sufficient to determine what a matched string represents. We therefore require the candidate secret or identifier to appear near a keyword, label, or field name associated with the target category, and refer to such rules as context-aware regular expressions. The objective of this study is to quantify the precision gained by making semantic context a mandatory part of the matching condition and the resulting cost in recall. We make three contributions: (1) a deterministic anchor-constrained rule model compatible with production regex catalogs, (2) paired ablation on CredData [4] and ai4privacy [5] that isolates the anchor's contribution, and (3) an empirical characterization of when anchoring helps—principally for ambiguous value formats—and when it does not.

II. RELATED WORK

Prior work shows both the scale of secret leakage and the limits of format-only detection. Meli et al. [1] and Sinha et al. [6] focused on credentials with distinctive structures, whereas SecretBench and subsequent tool

evaluations report substantial false-positive rates in broader secret detection [7], [2]. In industry, such noise contributes to warning fatigue and detector bypass [3].

False-positive reduction has also been approached with learned or hybrid methods, including lexical and contextual classifiers [8], PassFinder [9], and LLM-based post-filtering [10]. In PII detection, Presidio adjusts confidence using nearby context [11], CASSED combines headers and sample values [12], and more recent work likewise exploits surrounding text [13], [14], [15]. These approaches confirm the value of context but typically apply it as a score or in a post-processing stage.

Our formulation differs in three ways. Context is a mandatory conjunct: a value without a nearby anchor is not reported. The constraint is encoded directly in the regex and can therefore be shipped in existing YAML rule catalogs [16] without model inference or a separate filtering stage. Finally, paired ablation holds the value pattern fixed and removes only the anchor, allowing its precision-recall effect to be measured directly. Industrial scanners such as Gitleaks and TruffleHog also use keyword hints [17], [18], providing a practical point of comparison.

III. CONTEXT-AWARE RULE MODEL

Consider card_id: 5555-5555-5555-4444 and user_id: 5555-5555-5555-4444: a format-only card regex can match both, although only the first label supplies card semantics. We therefore model a rule as a value pattern constrained by a nearby context anchor, and report a value only when that anchor occurs within a bounded vicinity.

More precisely, a rule can be written as a tuple r = (V, δ_{pre}, δ_{post}), where A is a finite anchor vocabulary for the semantic category, V is the value pattern, and δ_{pre} and δ_{post} bound the distance, measured in characters or in lines, at which an anchor may precede or follow the value. A candidate string matching V is reported if and only if some anchor in A occurs within δ_{pre} units before it or within δ_{post} units after it. Prefix-anchored rules are the special case δ_{post} = 0; suffix anchors, as in a trailing comment such as "# customer card number" written after the value, correspond to δ_{pre} = 0; bidirectional windows are obtained by setting both bounds. The unit delta is fixed per rule: single-line rules, such as the canonical form above, bound characters, whereas the neighboring-line window analysis of Section VI bounds it in lines.

The following expression represents the prefix-anchored rule form:

(?i)(?:keyword_1|keyword_2|keyword_3) (?:[\s:=_-]{0,10}) (value_pattern)

The prefix form matches one of several labels, a bounded separator, and the value. The separator absorbs common assignment syntax while preventing the anchor from binding to an unrelated value farther along the line; suffix and bidirectional forms are obtained by reversing or alternating the order.

A card rule, for example, can require credit or card near a 13–19 digit candidate: (?i)\b(?: credit[\s_-]?)?card\w*[\s:=_-]{0,10}((?:\d[-]?){13,19}). Generic secrets benefit even more because their value patterns overlap heavily with those of ordinary random strings; anchors such as api_key, secret, client_secret, or authorization supply the missing semantic cue.

Accordingly, anchoring is expected to help most with ambiguous credentials and financial identifiers, less with personal data, and little with formats that are already intrinsically distinctive. Section VI tests this prediction.

IV. DESIGN AND IMPLEMENTATION

We keep ambiguous value patterns narrow and recover recall by expanding anchor vocabularies. Anchors are stored separately so that a value rule can be reused across languages; corpus-dependent validation, such as Luhn or mod-97, is applied outside the regex. The resulting catalog is optimized for triage quality rather than maximum detection volume.

We evaluate two implementations: a single regex that embeds anchor, separator, and value, and a two-stage variant that first locates anchors and then scans a bounded window for the value. The single-regex form is directly portable to production YAML catalogs, whereas the two-stage form permits a more expensive second pass. Runtime is compared in Section VI.

V. EMPIRICAL EVALUATION

1. Dataset and Setup

CredData [4] provides labeled suspicious lines for hard-coded credentials. We use 40 repositories selected by labeled-line count, yielding 48,856 labeled lines across 4,539 files and 8,668 targeted true findings. The paired ablation compares each context-aware rule with the same value pattern after removing the anchor; precision is measured at the line level and recall at the finding level. Format-intrinsic rules, for which anchor removal is not meaningful, serve as controls.

Anchor vocabularies were derived from field experience and common configuration-naming conventions, and were fixed before measurement; no English term was changed in response to CredData scores. Turkish password aliases produced no detections on this English corpus and were dropped, so the multilingual claim is evaluated on ai4privacy. Scripts, rules, and raw outputs are released in the replication package.

2. Results

Table 1. Context-aware vs format-only (anchor-ablated) rules on the CredData subset.

Metric	Context-aware	Format-only
Detected lines	7,235	5,649,099
True-positive lines	3,105	17,325
Overall precision	0.475	0.003
Recall (true findings)	0.356	0.956
Lines per true positive	2.3	326.1

Table 2. Per-category paired results: context-aware (CA) rule vs the same value pattern with the anchor removed (FO).

Category	CA det.	CA prec.	FO det.	FO prec.	Gain

Category	CA det.	CA prec.	FO det.	FO prec.	Gain
API key	48	0.432	1,527,044	0.0013	330×
Auth/access token	361	0.208	1,527,044	0.0013	159×
Token	364	0.202	1,527,044	0.0013	155×
Secret	279	0.542	1,527,044	0.0013	415×
Password	3,489	0.310	3,989,034	0.0007	453×
AWS secret key	0	—	9,980	0.0209	—
GitHub credential	41	0.244	64,212	0.0006	423×
Bearer/basic	1,277	0.563	3,185,996	0.0049	115×

Table 1 shows the main effect: anchoring raises overall precision from 0.003 to 0.475 while reducing candidate volume from 5.65 million to 7,235 lines. Table 2 shows large gains for every generic category with nonzero anchored matches (115×–453×); the AWS secret rule produced no anchored matches because benchmark obfuscation suppresses its exact assignment form. In contrast, intrinsically structured controls are already precise without anchors (PEM: 0.942, AWS key identifiers: 0.855, JWT: 0.769).

To account for repository-level dependence, we resampled the 40 repositories with replacement (B=5000). The 95% percentile interval is [0.282, 0.641] for anchored precision and [10, 631] for overall gain; despite the wide intervals, the per-category lower bounds for password and secret remain well above one.

3. Anchor Proximity and Window Size

Table 3. Relaxing the proximity constraint. Upper block: anchor window in lines over the generic categories. Lower block (italic): structured separator versus a pure character-distance bound on the assignment-shaped subset, which has its own baseline and is not comparable with the rows above.

Variant	Detections	Precision	Recall
Anchor adjacent (canonical)	7,235	0.475	0.356
Same line (w=0)	7,143	0.202	0.154
Within ±2 lines	42,605	0.036	0.165
Within ±5 lines	60,692	0.028	0.173
No anchor (ablated)	5,649,099	0.003	0.956
Structured separator	5,783	0.367	0.491
Pure distance δ=3	14,068	0.100	0.350
Pure distance δ=10	26,259	0.065	0.420
Pure distance δ=40	58,159	0.030	0.437

Table 3 shows that wider proximity windows quickly sacrifice precision while recovering little recall. Same-line matching falls to 0.202 precision, and extending the window to ±5 lines reaches only 0.173 recall at 0.028 precision. This suggests that expanding anchor vocabulary is a more promising recall lever than relaxing proximity.

4. Separator Model Ablation

The same table also tests whether the structured assignment separator contributes beyond what raw character distance alone provides. Replacing it with a pure distance bound reduces precision from 0.367 to at most 0.100 and does not recover the baseline recall. The separator, therefore, captures a syntactic anchor-value relationship rather than distance alone; this ablation applies only to assignment-shaped rules.

5. Generalization to Personal and Financial Data

To test generalization beyond credentials, we repeat the paired ablation on ai4privacy pii-masking-200k [5], 209,261 synthetic records in English, French, German, and Italian with span labels. The target categories deliberately include format collisions: the anchored variant requires a category keyword within 40 characters of the value, while the format-only variant uses the same value pattern.

Table 4. Context-aware vs format-only rules on the ai4privacy personal and financial corpus.

Category	FO prec.	CA prec.	Gain	FO rec.	CA rec.
Credit card	0.340	0.814	2.4×	1.000	0.303
Account number	0.270	0.738	2.7×	1.000	0.706
Phone	0.113	0.173	1.5×	0.946	0.093
SSN	1.000	1.000	1.0×	0.251	0.108
IBAN	0.982	1.000	1.0×	1.000	0.286

Table 4 reproduces the main pattern: anchoring more than doubles precision for the format-colliding credit-card and account-number categories. At the same time, SSN and IBAN gain no precision because their formats are already distinctive. The gains are consistent across all four languages, supporting the multilingual-anchor design.

Figure 1 explains the card result: 98% of format-only false positives are other PII with the same shape, mainly masked numbers, IMEIs, and IBANs. Format alone, therefore, cannot resolve the semantic category. Check-digit validation was not used because the synthetic corpus does not consistently preserve it; for example, only 10.3% of synthetic cards are Luhn-valid.

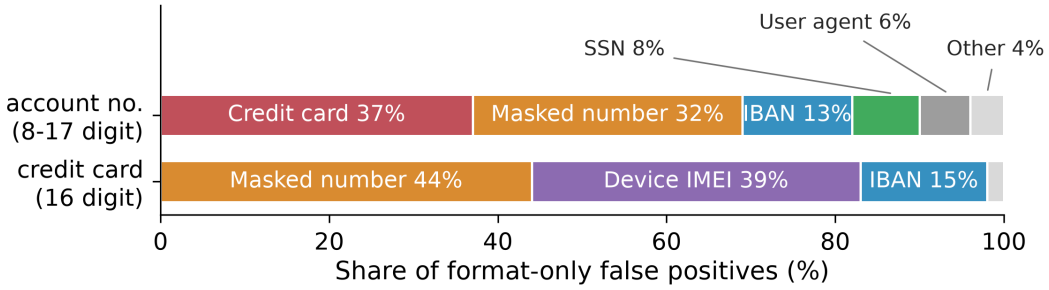


Figure 1. Composition of the format-only false positives for the credit-card and account-number rules on ai4privacy. Slices below 10% are labeled directly.

6. Comparison with Production Scanners

Existing scanners also use contextual keywords, so gitleaks 8.18.4, TruffleHog 3.82.6, detect-secrets 1.5.0, and Microsoft Presidio were run over the same CredData subset with their default rulesets and scored against the same ground truth. Results are reported at the line level and, where applicable, at the file level; file-level scoring is the fair unit for TruffleHog, whose filesystem mode reports no line numbers.

Table 5. Context-aware rules versus production scanners on the CredData subset.

Method	Det.	Line P	Line R	File P	File R
Context-aware (ours)	7,235	0.475	0.356	0.561	0.618
gitleaks	6,621	0.748	0.568	0.904	0.336
detect-secrets	8,591	0.261	0.174	0.526	0.594
TruffleHog	179	0.335	0.004	0.554	0.044
Format-only (ablated)	5,649,099	0.003	0.956	0.345	0.987

Without tuning thresholds or allowlists, our rules outperform detect-secrets and TruffleHog in line precision, while gitleaks is more precise (0.748 vs. 0.475). Most gitleaks findings come from its generic-api-key rule, an anchor-and-proximity construction augmented by entropy and a curated allowlist; the bare regex reaches 0.314 precision, entropy 0.443, and the allowlist 0.739. These post-filters are complementary to our rule model but were excluded because they would confound the anchor ablation.

7. Discussion and Limitations

Runtime was not a limiting factor: the six anchored assignment rules scan the 860 MB subset in 3.5 s (about 247 MB/s) on a commodity laptop. The two-stage implementation is slower (8.6 s) because the scanner already extracts literal keywords and efficiently skips non-matching lines; a separate prefilter is useful only when the second stage is genuinely expensive.

The central trade-off is recall. Anchored rules recover 35.6% of true findings, versus 95.6% for the ablated variants, so we recommend using the two modes at different operating points: anchored rules for high-confidence real-time alerts or blocking, and format-only rules for lower-priority exhaustive sweeps.

The evaluation has four main limitations. First, format-only variants are ablations of our rules rather than native production-scanner rules, so the paired comparison isolates the anchor contribution rather than tool superiority. Second, the repository subset is non-random, although bootstrap intervals quantify repository dependence. Third, CredData labels only examined lines; manual inspection of 100 detections on unlabeled lines found no plausible missed secret, limiting the possible precision shift to 0.475–0.479. Finally, the method assumes a nearby semantic label and therefore misses bare secrets, free-form cases, and unforeseen

anchor vocabulary. The canonical expressions avoid nested quantifiers and showed no pathological backtracking in our tests, but deployments on other regex engines should still validate rules and enforce per-match timeouts.

VI. CONCLUSION AND FUTURE WORK

Context-aware regular expressions provide a simple, deterministic way to trade recall for substantially higher precision when value formats are ambiguous. Across two public benchmarks, the benefit is strongest for generic credentials and colliding personal or financial identifiers, while intrinsically distinctive formats gain little. The method is therefore best used as a high-confidence layer alongside, not instead of, a high-recall sweep.

Future work will extend evaluation to the full CredData benchmark and additional country-specific identifiers, expand or learn anchor vocabularies automatically, and combine anchoring with orthogonal filters such as entropy and allowlists.

ACKNOWLEDGEMENT

The authors thank Trendyol and the Security Development Hub team for their support, and the maintainers of CredData and ai4privacy for releasing the benchmarks that made this evaluation possible.

REFERENCES

[1] M. Meli, M. R. McNiece, and B. Reaves, “How Bad Can It Git? Characterizing Secret Leakage in Public GitHub Repositories,” in Proc. Network and Distributed System Security Symp. (NDSS), San Diego, CA, USA, 2019.

[2] S. K. Basak, J. Cox, B. Reaves, and L. Williams, “A Comparative Study of Software Secrets Reporting by Secret Detection Tools,” in Proc. ACM/IEEE Int. Symp. Empirical Software Engineering and Measurement (ESEM), 2023.

[3] M. R. Rahman, N. Imtiaz, M.-A. Story, and L. Williams, “Why secret detection tools are not enough: It’s not just about false positives – An industrial case study,” Empirical Software Engineering, vol. 27, no. 3, art. 59, 2022.

[4] J. Yun, S. Choi, Y. Lee, O. Sokol, W. Shim, A. Melkonyan, and D. Kuzmenko, “Project CredData: A Dataset of Credentials for Research,” GitHub repository, 2021. [Online]. Available: https://github.com/Samsung/CredData

[5] ai4privacy, “pii-masking-200k: A Dataset for Removing Personally Identifiable Information from Text,” Hugging Face dataset, 2023. [Online]. Available: https://huggingface.co/datasets/ai4privacy/pii-masking-200k

[6] V. S. Sinha, D. Saha, P. Dhoolia, R. Padhye, and S. Mani, “Detecting and Mitigating Secret-Key Leaks in Source Code Repositories,” in Proc. IEEE/ACM 12th Working Conf. Mining Software Repositories (MSR), 2015, pp. 396–400.

[7] S. K. Basak, L. Neil, B. Reaves, and L. Williams, “SecretBench: A Dataset of Software Secrets,” in Proc. IEEE/ACM 20th Int. Conf. Mining Software Repositories (MSR), 2023.

[8] A. Saha, T. Denning, V. Srikumar, and S. K. Kasera, “Secrets in Source Code: Reducing False Positives Using Machine Learning,” in Proc. Int. Conf. Communication Systems & Networks (COMSNETS), 2020, pp. 168–175.

[9] R. Feng, Z. Yan, S. Peng, and Y. Zhang, “Automated Detection of Password Leakage from Public GitHub Repositories,” in Proc. IEEE/ACM 44th Int. Conf. Software Engineering (ICSE), 2022.

[10] M. N. Rahman, S. Ahmed, Z. Wahab, S. M. Sohan, and R. Shahriyar, "Secret Breach Detection in Source Code with Large Language Models," in Proc. ACM/IEEE Int. Symp. Empirical Software Engineering and Measurement (ESEM), 2025.

[11] Microsoft, "Presidio: Data Protection and De-identification SDK," 2024. [Online]. Available: https://microsoft.github.io/presidio/

[12] V. Kužina, A.-M. Petric, M. Barišić, and A. Jović, "CASSED: Context-based Approach for Structured Sensitive Data Detection," Expert Systems with Applications, vol. 223, art. 119924, 2023.

[13] L. Telkamp and M. Hulsebos, "Towards Contextual Sensitive Data Detection," arXiv:2512.04120 [cs.CR], 2025. (Preprint; no peer-reviewed version available at the time of writing.)

[14] M. Ponomarenko, S. Abedini, M. Shafieinejad, D. B. Emerson, S. Mohapatra, and X. He, "CAPID: Context-Aware PII Detection for Question-Answering Systems," in Proc. Student Research Workshop, Conf. European Chapter of the ACL (EACL), 2026.

[15] Z. Huang, K. Li, J. Yang, Y. Liu, and L. Wei, "Checked-In Secret Detection: Strings Are All You Need," in Proc. ACM SIGSOFT Int. Symp. Software Testing and Analysis (ISSTA), 2026, doi: 10.1145/3832207.

[16] Trendyol, "Awesome Regex List," GitHub repository, 2024. [Online]. Available: https://github.com/Trendyol/awesome-regex-list

[17] Z. Rice, "Gitleaks," GitHub repository, 2024. [Online]. Available: https://github.com/gitleaks/gitleaks

[18] Truffle Security, "TruffleHog," GitHub repository, 2024. [Online]. Available: https://github.com/trufflesecurity/trufflehog

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Privacy-Preserving Secret Validation: An Empirical Comparison of Probabilistic Filters, k-Anonymity, and Private Set Intersection

Burak Tahtacı
Security Development Hub
Trendyol Group
Istanbul, Türkiye
burak.tahtaci@trendyol.com

Enes Can Güven
Security Development Hub
Trendyol Group
Istanbul, Türkiye
enescan.guven@trendyol.com

Tuna Çakar
Computer Engineering Dept.
MEF University
Istanbul, Türkiye
cakart@mef.edu.tr

Abstract
Systems increasingly need to determine whether a given secret has been seen before without storing that secret in plaintext. This is the problem underlying breached-password detection, credential-stuffing defense, and duplicate-secret prevention. Individual protocols have been analyzed in isolation, but their comparative behavior under identical experimental conditions is not well documented. We empirically compare three families: probabilistic membership filters (Bloom and Cuckoo), k-anonymity hash-prefix lookups, and Private Set Intersection and Oblivious Pseudorandom Function protocols, benchmarking both a classic Diffie-Hellman OPRF and a modern Ristretto255 OPRF. Under a threat model that separates stored-set privacy, query privacy, and offline brute-force resistance, and using a directly measured dataset of 100,000 secrets, we evaluate build cost, memory, latency, communication, error rate, and privacy. Filters answer in one to two microseconds with a 117–160KB footprint but leak membership to an offline attacker; k-anonymity adds query privacy for a thirty-fold larger stored set and 35 bytes of online communication at identical latency; the DH-OPRF removes offline brute-force leakage at 514 bytes online and about 104ms per query, while the Ristretto255 OPRF achieves the same guarantees with 32-byte storage and millisecond-scale queries. We distill these trade-offs into a decision guide and release our implementations and benchmark harness as open source.

Keywords: privacy, bloom filter, cuckoo filter, k-anonymity, private set intersection, oblivious pseudorandom function, breached credentials, membership testing

I. INTRODUCTION

A growing number of systems must check whether a user-supplied secret, such as a password, API token, recovery code, or cryptographic key, appears in some reference set, without learning or storing the secret itself. Use cases include breached-credential detection, secret scanning in committed files or logs, and duplicate or weak-secret prevention. Have I Been Pwned (HIBP) and Google Password Checkup are production deployments of the first case. The naive solution, storing the secrets and comparing, violates the very property these systems exist to protect. A range of techniques sidestep plaintext storage, but they are frequently conflated under the banner of "privacy-preserving" despite offering materially different guarantees: a Bloom filter and a Private Set Intersection protocol both avoid storing plaintext, yet only one resists an attacker who steals the server state and brute-forces, low-entropy candidates. Our contributions are:

- A threat model (Section III) decomposing privacy into stored-set privacy, query privacy, and offline brute-force resistance.
- Open implementations (Section V) of Bloom and Cuckoo filters, k-anonymity hash-prefix lookup, and a DH-based PSI/OPRF protocol.

- An empirical comparison (Section VI) on a shared 100K-secret dataset across build cost, memory, latency, communication, and false-positive rate.
- A decision guide (Table 4) mapping deployment requirements to a recommended method.

II. BACKGROUND AND RELATED WORK

1. Probabilistic membership filters

A Bloom filter [1] stores a set in a bit array of m bits using k hash functions; membership tests have no false negatives but a tunable false-positive rate. Bonomi et al. [2] extended this model with Counting Bloom Filters that support item deletion. Fan et al. [3] introduced Cuckoo filters, which store short fingerprints in a cuckoo-hashed table, offering deletion support and a better space–error trade-off at low target rates. Graf and Lemire [4] further optimized space and retrieval speed with XOR and Binary Fuse filters. Naor and Yogev [5] showed that standard filters can leak set members when queried adaptively by an adversary. All of these filters are compact and local: a query touches only the structure, with no network communication.

2. K-anonymity hash-prefix lookup

Hunt [6] popularized the k-anonymity prefix lookup with the Have I Been Pwned Pwned Passwords API: the client hashes its secret, sends only a short prefix of the digest, and receives every stored digest sharing that prefix. The client completes the comparison locally, so the server never sees the secret itself, only a prefix that is shared by an anonymity set of many candidate digests. Li et al. [7], [8] rigorously analyzed the security and privacy bounds of various bucketization techniques for this protocol. Golla et al. [9] showed in a large-scale field study that speed of validation is critical to real-world adoption.

3. Private Set Intersection and Oblivious PRF

Private Set Intersection (PSI) lets two parties learn the intersection of their sets without revealing the remainder, dating back to Meadows [10] and scaled via oblivious transfer extensions by Pinkas et al. [11]. Modern systems reduce per-query cost with Oblivious Pseudo-Random Functions (OPRFs), introduced by Kolesnikov et al. [12] and standardized for prime-order groups in RFC 9497 by Davidson et al. [13]. Google Password Checkup [14] provides the first massive-scale deployment: the client blinds its secret, the server applies a secret key, and the client unblinds to obtain a pseudorandom value comparable against a bucketed encrypted set. The server never sees the secret, and the stored pseudorandom group elements are not offline brute-forceable without the key. Angelou et al. [15] extended OPRF-based membership tests with asymmetric protocols that resist malicious servers, and Bellare et al. [16] and Krawczyk et al. [17] showed how HMAC-style keyed hashing prevents offline dictionary attacks against keyed checks.

4. Comparative Studies

Archer et al. [18] survey real-world deployments of secure computation, with private membership checks as one application, and Ateniese et al. [19] analyze proofs of storage with public verifiability. Yet, both focus on a single protocol family. The works surveyed above each propose, optimize, or deploy an individual primitive; none conducts a side-by-side empirical comparison of probabilistic filters, k-anonymity, and PSI/OPRF under a shared threat model, dataset, and benchmarking environment. This paper aims to fill that gap.

III. THREAT MODEL AND PRIVACY DEFINITIONS

We consider a client holding a secret x *and a server holding* a set büyük harf S.We evaluate each method against three independent adversarial goals. A method may satisfy any subset.

(P1) Stored-set privacy. An attacker who obtains the server's persisted state (filter bits, digest list, or encrypted set) should not be able to recover the secrets in S. We further distinguish offline brute-force resistance: even given the state and a dictionary of candidate secrets, can the attacker confirm which ccandidates arein S.?
(P2) Query privacy. The server, or a network observer, should not learn the queried secret x, beyond an explicitly bounded leakage such as a hash prefix defining an anonymity set of size k.
(P3) Correctness and no false negatives. A member must never be reported absent. False positives are tolerated at a tunable rate for filters and must be accounted for in the threat analysis: a false positive can cause an unnecessary rejection, but never a missed match.

Table 1. Qualitative privacy and communication placement of each method.

Method	P1 set	Offline BF	P2 query	Offline Setup/Dist.	Online Query Comm.
Bloom filter	partial	leaks	local only	High (Full index download)	0 B (local query)
Cuckoo filter	partial	leaks	local only	High (Full index download)	0 B (local query)
k-anonymity	digest list	leaks	k-anon set	0 B (hosted on server)	Small (35 B per query)
PSI / OPRF	pseudorandom	resists	blinded	0 B (hosted on server)	Moderate (514 B per query)

Table 1 encodes a central insight: *compactness is not privacy*. Filters and k-anonymity avoid plaintext storage yet leak set membership to an offline attacker; only PSI and OPRF close that gap, and Section VI measures the cost. The table also highlights an architectural trade-off in network overhead: filters require a large one-time offline download but query locally with zero bytes over the network, while k-anonymity and PSI host the reference set server-side but incur network overhead on every query.

IV. EXPERIMENTAL DESIGN

We evaluate on two distinct corpora designed to capture the full spectrum of entropy conditions encountered in practice: Random high-entropy: one hundred thousand random alphanumeric strings of length 8–64 (Python's secrets; scripts/generate_secrets.py), representing API tokens, recovery codes, and salts with near-maximal entropy. RockYou low-entropy: one hundred thousand unique printable ASCII passwords of length 4–64 from the RockYou breach corpus (scripts/extract_rockyou.py), filtered to match the random dataset cardinality. These are human-chosen secrets with severe repetition, low entropy, and a heavy-tailed Zipfian distribution. The two datasets span opposite entropy extremes; Section VI re-runs the full benchmark on RockYou to expose how each method behaves under low-entropy, skewed distributions.

Probabilistic Filters (Bloom & Cuckoo):

Empirical result: The measured FP rate (over 100,000 never-seen items) is 0.99% vs 1.04% (Bloom) and 0.60% vs 0.60% (Cuckoo) for random vs RockYou, a difference well inside the 95% confidence interval. The probabilistic error bounds hold under both entropy conditions (Section VI).

Security Vulnerability: Despite stable FP rates, low-entropy filters are highly vulnerable to offline brute-force. An attacker with the filter can test a local RockYou dictionary in microseconds to reveal membership in the set.

k-Anonymity Prefix Lookup:

Empirical result: Both datasets yield a median anonymity set of 1 and a maximum bucket size of 4 at a 20-bit pre10, son taban, üssü altıncı, bucket sparsity dominates regardless of the entropy distribution (see anonymity-set caveat).

Scale Mitigation: IHIBP ($N \approx 10^9$), the density parameter $\lambda = \frac{N}{B} \approx$ 954 guguarantees anonymity , underweightedistributions.

Cryptographic PSI and OPRF:

Empirical result: PSI build time is 45.8 min vs 45.8 min and query latency 103.6 ms vs 105.0 ms (about 1.3% difference) on the full 100K dataset. The two corpora are functionally identical, confirming that query cost is independent of secret entropy.

V. IMPLEMENTATION

Bloom. The bit count m and hash count kk follow the standard optimal formula in (1), with MurmurHash3 and incrementing seeds over a bit-packed array:

$$m = -\frac{nln\varepsilon}{(ln2)^2}, \qquad k = \frac{m}{n} ln2. \tag{1}$$

Luck. Four-slot buckets, fingerprint size $f = \lceil \log_2(2b/\varepsilon) \rceil$or bucket size b, a power-of-two table for XOR-based alternate indexing, and up to five hundred eviction kicks.

k-anonymity. SHA-1 digests bucketed by a five-hex-character (twenty-bit) prefix, matching HIBP. The client sends the prefix and compares the returned suffix digests locally.

PSI and OPRF. Our baseline uses a Decisional Diffie-Hellman (DH) construction over the RFC 3526 2048-bit MODP group, modeling the corof Google Password Checkup [14]: with server key k the stored value is $F_k(y) = g^{H(y)\cdot k}$; the client blinds with random r, the server raises to k, and the client unblinds with r^{-1}. The encrypted set is bucketed by a sixteen-bit prefix for k-anonymity-style scaling (256 bytes/item stored; 514 bytes online per query, see Section VI). A modern baseline uses the same OPRF structure over the Ristretto255 group (RFC 9496 [20]) following RFC 9497 [13], with 32-byte elements and no dependence on the legacy 2048-bit group. Both variants are pure Python with no external cryptographic dependencies, relying only on a hash function and modular exponentiation; mature open-source Ristretto implementations exist [21], so our pure-Python EC-OPRF latency bound is conservative.

VI. RESULTS

Table 2 reports all metrics for 100,000 secrets; every method, including the DH-based PSI's 45-minute public-key build, was measured directly on the complete dataset, so no values are extrapolated. False-positive rates are estimated over 100,000 never-seen items rather than the 2,000 used for latency, giving a narrow 95% confidence interval.

Table 2. Benchmark results (100K random secrets).

Metric	Bloom	Cuckoo	k-anon	PSI/OPRF	EC-OPRF
Build time (100k)	0.17 s	0.09 s	0.12 s	45.8 min	4.4 min
Stored size (100k)	117 KB	160 KB	3.4 MB	25.6 MB	3.2 MB
Stored size (per item)	1.2 B	1.6 B	35 B	256 B	32 B
Latency present (μs)	1.58	0.88	1.33	103,632	5,416
Latency absent (μs)	0.58	0.96	1.04	103,550	5,321
Offline distribution cost	117 KB	160 KB	0 B	0 B	0 B
Online query communication	0 B	0 B	35 B	514 B	96 B
False-positive rate	0.99%	0.60%	0%	0%	0%
Median anonymity set	—	—	1	1	2

Re-running the full benchmark on the RockYou corpus (Table 3) leaves every cross-method relative ranking unchanged: build times, stored sizes, and latencies move by at most a few percent, with Bloom's false-positive rate shifting from 0.99% to 1.04%. The privacy and cost ladder of Section III is thus structurally robust to data entropy.

Table 3. Low-entropy (RockYou) results.

Metric	Bloom	Cuckoo	k-anon	PSI/OPRF	EC-OPRF
Build time (100k)	0.16 s	0.08 s	0.12 s	45.8 min	4.5 min
Stored size (100k)	117 KB	160 KB	3.4 MB	25.6 MB	3.2 MB
Latency present (μs)	1.54	0.83	1.33	105,028	5,458
FP rate	1.04%	0.60%	0%	0%	0%
Median anon. set	—	—	1	1	2

Latency. As shown in Figure 1, probabilistic filters and k-anonymity all answer in one to two microseconds. The DH-based PSI answers in about one hundred milliseconds (103,632 μs), roughly sixty-six thousand times slower than a filter, because each query performs three 2048-bit modular exponentiations for blinding, evaluation, and unblinding; the EC-OPRF answers in 5,416 μs (pure Python), and a compiled curve25519/ristretto backend would place it in the 0.3–3 ms range. This is the price of cryptographic stored-set privacy.

Filter and k-anonymity latencies measure only local computation. A deployed k-anonymity or PSI/OPRF query additionally incurs a network round-trip: 0.47 ms median measured on loopback, 10–100 ms on a wide-area network. Adding one loopback RTT to the server-held methods collapses the k-anonymity-versus-DH-OPRF gap (Figure 1, solid markers); over a wide-area RTT, both sit within a constant factor because the

network leg dominates. Stored size. Filters are the most compact, under two bytes per item (117–160 KB at full scale), followed by k-anonymity (35 B/item, 3.4 MB), the EC-OPRF (32 B/item, 3.2 MB), and the DH-OPRF, whose 256-byte group elements total 25.6 MB, the price of brute-force-resistant pseudorandom storage. In all cases, per-item storage is smaller than the storage required for online query communication, which additionally carries the request element and the downloaded bucket. False positives. Both probabilistic filters track their one percent target: 0.99% (Bloom, 95% CI 0.93–1.05%) and 0.60% (Cuckoo, 95% CI 0.56–0.65%) over the 100,000-item never-seen sample. The hash-based methods are exact, with zero false positives, since they compare full digests or group elements rather than short probabilistic fingerprints. Communication. Probabilistic filters require a large one-time offline download (117–160 KB) but query locally with zero bytes online. k-anonymity and OPRFs host the reference set server-side (zero offline cost) but incur per-query costs: 35 bytes for k-anonymity, 514 bytes for the DH-OPRF, and 96 bytes for the EC-OPRF (Section V). Privacy versus cost. Figure 2 plots query latency against the privacy ladder of Section III: filters sit at the cheap, low-privacy corner; k-anonymity adds query privacy at near-zero extra latency; the DH-OPRF occupies the high-privacy corner roughly five orders of magnitude to the right; and the EC-OPRF shortens that gap to about 3.5 orders of magnitude.

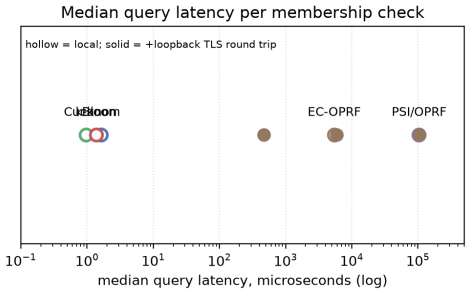


Figure 1. Median query latency per membership check, log scale.

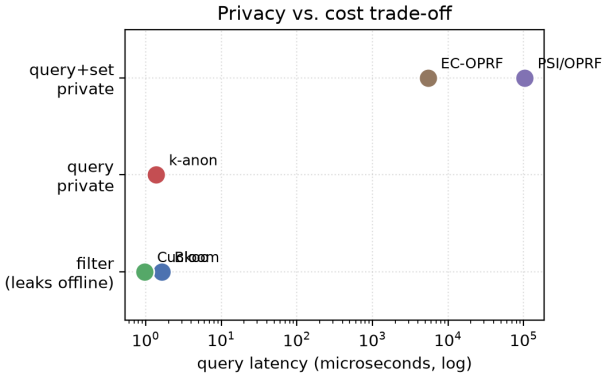


Figure 2. Privacy versus cost.

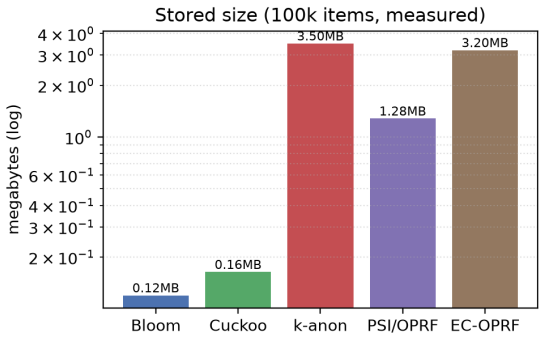


Figure 3. Stored size per method, measured (not extrapolated) on one hundred thousand items, log scale. EC-OPRF shows the 32-byte-element advantage over the DH group.

1. The anonymity-set caveat

Our high-entropy dataset yields a median bucket size of 1 for the hash-prefix method with a 20-bit prefix (and 2 for the OPRF variants, which use a shorter 16-bit prefix). In this corpus, a prefix often uniquely identifies the secret, providing little practical query privacy. Distributing N secrets across B independent prefix buckets, with SHA-1 behaving as a random oracle, gives a Poisson process with mean density per bucket:

$$\lambda = \frac{N}{B} \tag{2}$$

In our setup, $N = 100{,}000$ and a 5-hex-character (20-bit) prefix yields $B = 2^{20} = 1{,}048, buckets$, so the density parameter is:

$$\lambda = \frac{100{,}000}{1{,}048{,}576} \approx 0.0954 \tag{3}$$

The Poisson distribution gives the probability of exactly $ksecrets$ in a bucket and, from (3), the probability that a bucket is non-empty:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad P(X \geq 1) = 1 - e^{-\lambda} \approx 0.0910. \tag{4}$$

Conditioned on a bucket being active, the probability that it contains exactly one secret is therefore

$$P(X = 1 \mid X \geq 1) = \frac{P(X = 1)}{1 - P(X = 0)} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} \approx 0.953, \quad (5)$$

ov 95% of, a, activ, bucket, contai, extly, ne, sect,, which expinshyhenonymity set is one.

Now consider a massive dataset like Have I Been Pwned (HIBP), where N is approximately equal to 10^9; secrets are indexed using the same 5-hex prefix (B = 1,048,576) The density parameter rises to:

$$\lambda = \frac{10^9}{1{,}048{,}576} \approx 953.7 \quad (6)$$

Because λ is large, the Poisson distribution can be approxmated by a Gaussi an distribution $N(\mu = 953.7, \sigma^2 = 953.7)$, with a standard deviation of:

$$\sigma = \sqrt{953.7} \approx 30.9 \quad (7)$$

The central takeaway is that k-anonymity's query privacy is scale-dependent: it is a direct function of the density ratio, büy harf B $B.A$ small or sparse set degrades to near-plaintext prefixes; we report this artifact rather than tuning the dataset to flatter the method. Prefix-length sensitivity. The prefix length trades anonymity-set size against download cost. On the 100K dataset, the 20-bit prefix (five hex characters, HIBP's choice) yields the median bucket of one that was already reported. A 16-bit prefix ($B = 65{,}536$) raises the density to $\lambda \approx 1.53$ and the median bucket to two (maximum nine, the EC-OPRF configuration in Table 2), at the cost of roughly twice the suffix data per query. Shortening to 12 bits ($(B = 4{,}096, lambda \approx 24)$ $restores\ a\ median\ bucket\ of\ around$ 20Still, it downloadsmost of the set for a single query, defeating the purpose. We therefore recommend adaptive prefix selection: choose B so that $\lambda = N/B$ exceeds the desired anonymity bound (subject to the response-size budget. Our PSI/OPRF implementations expose this as a configurable parameter.

Table 4. Decision guide: mapping deployment priorities.

Priority	Set	Use
Smallest footprint, fast local check	public or shippable	Bloom or Cuckoo
Same, plus deletions	public or shippable	Cuckoo
Hide the query from a server	large, server-held	k-anonymity
Hide query and stored set; offline resistance	sensitive, server-held	EC-OPRF preferred
Exact membership, modest privacy; low latency	server-held	k-anonymity
Match existing Checkup-style deployment	sensitive, server-held	DH-OPRF (legacy)



VII. CONCLUSION

Privacy-preserving secret validation is not one guarantee but a ladder of them. We decomposed it into stored-set privacy, query privacy, and offline brute-force resistance, and measured five method variants spanning the ladder on a common 100K-secret benchmark. The practical takeaway is that compactness and privacy are distinct properties with a steep exchange rate, and the right choice follows directly from which of the three guarantees a deployment actually needs.

REFERENCES

[1]. B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," Commun. ACM, vol. 13, no. 7, pp. 422–426, 1970.
[2]. F. Bonomi, M. Mitzenmacher, R. Panigrahy, S. Singh, and G. Varghese, "An improved construction for counting Bloom filters," in Proc. 14th Annual European Symp. Algorithms (ESA), 2006, pp. 684–695.
[3]. B. Fan, D. G. Andersen, M. Kaminsky, and M. D. Mitzenmacher, "Cuckoo filter: practically better than Bloom," in Proc. 10th ACM Int. Conf. Emerging Networking Experiments and Technologies (CoNEXT), 2014, pp. 75–88.
[4]. T. M. Graf and D. Lemire, "Xor filters: faster and smaller than Bloom and cuckoo filters," ACM J. Experimental Algorithmics, vol. 25, pp. 1–16, 2020.
[5]. M. Naor and G. Yogev, "Bloom filters in adversarial environments," in Proc. Annual Cryptology Conf. (CRYPTO), 2015, pp. 565–584.
[6]. T. Hunt, "I've just launched Pwned Passwords version 2," troyhunt.com, 2018. [Online].
[7]. L. Li, B. Pal, J. Ali, N. Sullivan, R. Chatterjee, and T. Ristenpart, "Protocols for checking compromised credentials," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), 2019, pp. 1387–1403.
[8]. L. Li, B. Pal, J. Ali, E. Fernandes, S. Chatterjee, and T. Ristenpart, "Securing compromised credentials checking in practice," ACM Trans. Privacy and Security (TOPS), vol. 27, no. 1, pp. 1–30, 2024.
[9]. M. Golla, M. Wei, J. Hainline, L. Filipe, M. Dürmuth, E. M. Redmiles, and B. Ur, "What was that site doing with my Facebook password? Designing password-reuse notifications," in Proc. IEEE Symp. Security and Privacy (S&P), 2018, pp. 1549–1566.
[10]. C. Meadows, "A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party," in Proc. IEEE Symp. Security and Privacy, 1986, pp. 134–137.
[11]. B. Pinkas, T. Schneider, and M. Zohner, "Scalable private set intersection based on OT extension," ACM Trans. Privacy and Security, vol. 21, no. 2, pp. 1–35, 2018.
[12]. V. Kolesnikov, R. Kumaresan, M. Rosulek, and N. Trieu, "Efficient batched oblivious PRF with applications to private set intersection," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), 2016, pp. 813–824.
[13]. A. Davidson, J. Ali, and N. Sullivan, "Oblivious Pseudorandom Functions (OPRFs) using prime-order groups," Internet Engineering Task Force (IETF), RFC 9497, 2023. [Online].
[14]. K. Thomas et al., "Protecting accounts from credential stuffing with password breach alerting," in Proc. 28th USENIX Security Symp., 2019, pp. 1556–1571.
[15]. N. Angelou et al., "Asymmetric private set intersection with applications to contact tracing and private vertical federated machine learning," arXiv:2011.09350, 2020. [Online].
[16]. M. Bellare, R. Canetti, and H. Krawczyk, "Keying hash functions for message authentication," in Proc. Annual Cryptology Conf. (CRYPTO), 1996, pp. 1–15.



[17]. H. Krawczyk, M. Bellare, and R. Canetti, "HMAC: Keyed-hashing for message authentication," Internet Engineering Task Force (IETF), RFC 2104, 1997. [Online].
[18]. D. W. Archer, D. Bogdanov, Y. Lindell, L. Kamm, K. Nielsen, J. Illeborg Pagter, N. P. Smart, and R. N. Wright, "From keys to databases—real-world applications of secure multi-party computation," The Computer Journal, vol. 61, no. 12, pp. 1749–1771, 2018.
[19]. G. Ateniese, S. Kamara, and J. Katz, "Proofs of storage from homomorphic identification protocols," in Proc. Advances in Cryptology (ASIACRYPT), 2009, pp. 319–333.
[20]. H. de Valence, J. Grigg, M. Hamburg, and I. Lovecruft, "Ristretto255: A group for the Ristretto group," Internet Research Task Force (IRTF), RFC 9496, 2023. [Online].
[21]. L. Quinn, F. Jansen, and K. Bab, "OPRF: Oblivious pseudorandom function over an elliptic curve (Ristretto255)," GitHub repository multiparty/oprf, 2018–2023. [Online].

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Evaluating the Impact of Data Augmentation on Cross-Dataset Generalization in Intrusion Detection Systems

Mounira Zekiouk
Department of Computer Science
University of Mila
MISC Laboratory, Constantine 2 University.
Mila, Algeria
m.zekiouk@centre-univ-mila.dz

Madjed Bencheikh Lehocine
Department of Computer Science,
University of Mila
LISI Laboratory, University of Mila
Mila, Algeria
m.bencheikh-lehocine@centre-univ-mila.dz

Ahlam Bouanane
Department of Computer Science
University of Mila, Algeria
bouananeahlam33@gmail.com

Abstract
Machine-learning-based intrusion detection systems (ML-based IDSs) often achieve high performance under intra-dataset evaluation but exhibit substantial degradation when applied to unseen datasets due to class imbalance and distributional differences. This study is based on two benchmark intrusion detection datasets, CIC-IDS2017 and CIC-IDS2018, and investigates whether data augmentation can improve the cross-dataset generalization of binary and multiclass intrusion detection models. Before model development, both datasets were cleaned by removing duplicate records and samples containing missing values. Their feature spaces were then harmonized by retaining common features, and attack labels were mapped to a unified taxonomy. Models were subsequently trained on the CIC-IDS2017 dataset using five data augmentation techniques and their corresponding non-augmented baselines. Their performance was assessed through intra-dataset evaluation on CIC-IDS2017 and cross-dataset evaluation on CIC-IDS2018. Without augmentation, multiclass macro recall decreased from 0.99 to 0.48 across the two evaluation settings, while binary Attack recall declined from 0.95 to 0.45. Among the evaluated techniques, SMOTE achieved the strongest cross-dataset performance, increasing multiclass macro recall to 0.88 and binary Attack recall to 0.78, followed by random oversampling with 0.84 and 0.70, respectively. GAN-based augmentation and ADASYN yielded moderate improvements, whereas MixUp was less effective. These findings indicate that appropriate data augmentation can substantially enhance cross-dataset intrusion detection, although additional approaches are needed to address distributional differences across heterogeneous network datasets better.

Keywords: Intrusion Detection System; Cross-Dataset Generalization; Data Augmentation; Class Imbalance; Machine Learning.

I. INTRODUCTION

As enterprise networks grow in scale and complexity, ML-based IDSs have become important for automatically detecting malicious network activities [1]. However, their real-world effectiveness is often limited by class imbalance, as benign traffic dominates while relatively few samples represent many critical attacks. This imbalance can bias models toward majority classes, resulting in high overall accuracy but poor detection of rare and security-critical attacks [2],[3]. Several data augmentation techniques have been proposed to address this problem [4],[8]. Although these methods often improve detection performance, most studies evaluate them only within the same dataset. Therefore, it remains unclear whether these improvements generalize to unseen network environments or mainly reflect adaptation to dataset-specific traffic characteristics. This issue is important because distributional differences between datasets can substantially reduce IDS performance [9],[10], while the impact of augmentation on cross-dataset generalization remains insufficiently explored [11].

Accordingly, this study evaluates multiple data augmentation techniques to improve cross-dataset generalization of binary- and multiclass ML-based IDS models. Models are trained on CIC-IDS2017 and evaluated in both intra- and cross-dataset settings, with CIC-IDS2018 as the external dataset. The study aims to determine whether augmentation helps models learn more transferable attack representations and maintain effective detection on previously unseen network traffic. The findings can benefit cybersecurity researchers, IDS developers, and network security practitioners by guiding the selection of augmentation techniques to enhance intrusion detection robustness. The study also demonstrates the importance of cross-dataset evaluation when assessing IDS generalization and provides a basis for further research on domain adaptation, distribution alignment, and multi-dataset learning. The remainder of this paper is organized as follows. Section 2 reviews related work on ML-based IDSs, data augmentation, and cross-dataset generalization. Section 3 presents the datasets, preprocessing procedures, augmentation techniques, and evaluation methodology. Section 4 presents and discusses the experimental results, and Section 5 concludes the paper and outlines future research directions.

II. RELATED WORK

ML has become the dominant paradigm for Network Intrusion Detection Systems (NIDSs), progressively replacing traditional signature-based approaches with data-driven techniques. Over the past decade, extensive research has investigated the application of ML to intrusion detection, resulting in a wide variety of supervised, unsupervised, ensemble, and deep learning models. Recent comprehensive surveys [12-16] have highlighted the rapid evolution of ML-based IDSs, covering advances in learning algorithms, feature engineering, benchmark datasets, and evaluation methodologies. Despite these significant advances, the literature consistently identifies several unresolved challenges, including class imbalance, dataset bias, model robustness, explainability, and the limited ability of current models to generalize across heterogeneous network environments. Among these challenges, class imbalance is particularly important because minority attack classes are often underrepresented, which may considerably reduce their detection rates. To address this problem, various data augmentation techniques have been proposed to generate additional minority-class samples without discarding majority-class traffic [17]. In network intrusion detection, these techniques must preserve the statistical and behavioral characteristics of high-dimensional tabular network data to avoid generating unrealistic traffic samples [18]. Existing approaches can be broadly classified into three categories. Distance-based oversampling methods, such as SMOTE and ADASYN, generate new samples by interpolating between neighboring minority instances, but they may introduce noisy or ambiguous samples in sparse feature spaces [19], [20]. Deep generative models, including GANs, VAEs, and CTGANs, learn the underlying data distribution to generate more realistic synthetic traffic, although they often suffer from training instability and high computational complexity [18],[21],[22]. Feature-interpolation methods, such as MixUp, improve model regularization through convex combinations of training samples. Nevertheless, their effectiveness in tabular network traffic remains insufficiently explored, as feature interpolation may alter the semantic relationships among flow characteristics [23]. Although data augmentation is widely used to address class imbalance, its practical value also depends on whether it helps models learn patterns that generalize beyond the training dataset. Strong intra-dataset performance does not necessarily translate to robustness under different network conditions. Cross-dataset validation, where models are trained on one dataset and evaluated on another, provides a more realistic assessment of generalization [24],[25]. However, despite growing interest in both data augmentation and cross-dataset evaluation, the effectiveness of augmentation in improving cross-dataset generalization remains underexplored. Therefore, this study systematically evaluates five representative data-augmentation techniques—SMOTE, ADASYN, GANs, MixUp, and random oversampling (RO)—within a harmonized experimental framework based on the CIC-IDS2017 and CSE-CIC-IDS2018 datasets. Unlike most previous studies that focus primarily on intra-dataset performance, this work examines whether augmentation can improve the transferability of learned attack patterns across heterogeneous network environments. The

models are evaluated in both binary and multiclass settings, in intra- and cross-dataset scenarios, enabling a direct comparison of augmentation effectiveness under distribution shift. This analysis aims to clarify the extent to which data augmentation can mitigate the generalization limitations of ML-based IDSs.

III. METHODOLOGY & EXPERIMENTAL SETUP

This section presents the experimental framework adopted to assess the impact of data augmentation on cross-dataset intrusion detection. As illustrated in the figure. 1, the proposed methodology comprises five main steps: (1) dataset cleaning and feature harmonization; (2) dataset splitting; (3) training-set preparation, including categorical encoding, normalization, feature selection, and the application of the selected data augmentation techniques; (4) model training; and (5) performance evaluation under both intra-dataset and cross-dataset settings.

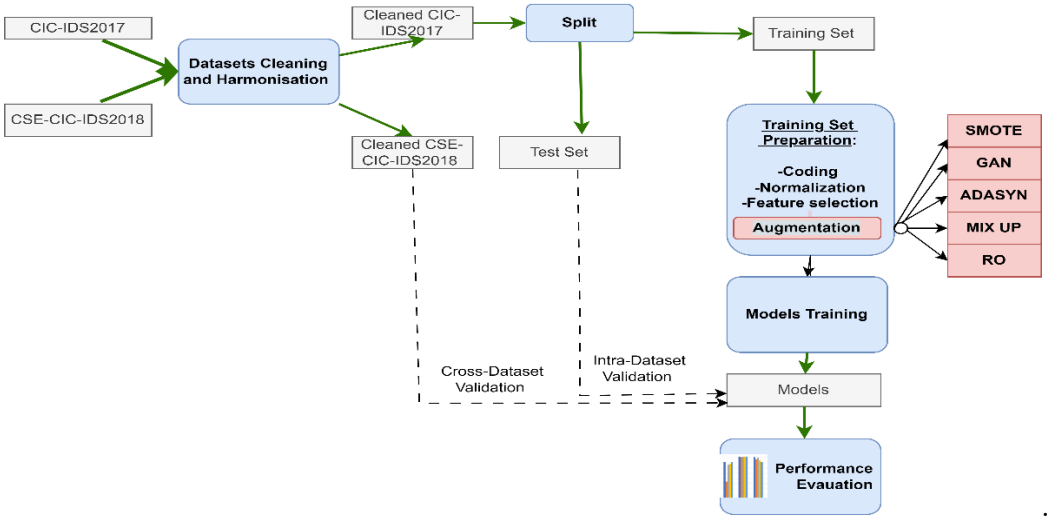


Figure 1. Overview of the proposed experimental framework.

1. Datasets

This study employs two publicly available benchmark intrusion detection datasets: CIC-IDS2017 and CSE-CIC-IDS2018 [26],[27]. CIC-IDS2017 was developed by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick [26], while CSE-CIC-IDS2018 was developed through a collaboration between CIC and the Communications Security Establishment (CSE) [27]. Both datasets contain labeled benign and malicious network traffic and are available in PCAP and CSV formats. Their flow-based features were extracted using CICFlowMeter, providing more than 80 network traffic features suitable for ML analysis. In this study, the CSV versions were used and harmonized to obtain a common feature representation for cross-dataset evaluation. Despite their related feature spaces, the datasets differ in traffic distributions, attack scenarios, and network environments, making them suitable for evaluating the generalization capability of ML-IDSs.

2. Datasets Cleaning and Harmonization

To ensure a consistent and reliable experimental framework, both CIC-IDS2017 and CSE-CIC-IDS2018 datasets underwent a comprehensive cleaning and harmonization process before model development. First, duplicate network flows were removed to eliminate redundant observations. Subsequently, samples containing missing, infinite, or invalid feature values were identified and discarded. After data cleaning, the feature schemas of both datasets were harmonized by retaining only features common to both, thereby establishing a unified feature space for cross-dataset evaluation. Finally, the original attack labels were reorganized into a common taxonomy by grouping semantically related attacks into seven unified categories: DoS**,** DDoS**,** Bot**,** Patator**,** Web Attack**,** PortScan**,** and DoS Hulk.

3. Dataset Splitting Strategy

To ensure a reliable and unbiased evaluation, the CIC-IDS2017 dataset was partitioned using a stratified splitting strategy that preserved the original class distribution. Specifically, 70% of the data were allocated to model training, 10% to validation, and the remaining 20% to final testing.

4. Training Set Preparation

- The training data were first encoded to transform categorical class labels into numerical representations suitable for ML. For the binary classification task, Label Encoding was applied by assigning 0 to benign traffic and 1 to attack traffic. For the multiclass classification task, One-Hot Encoding was employed to represent the different attack categories.
- Feature normalization was applied to reduce the effect of differences in feature scales and improve model stability. Two normalization techniques, Robust Scaler and Min-Max Scaling, were experimentally evaluated for both binary and multi-class classification.
- Feature selection was performed to reduce dimensionality and improve model performance. A Random Forest classifier was used to rank features by importance score. Multiple candidate subsets of the highest-ranked features were systematically evaluated, and the optimal subsets were selected based on the best validation performance, yielding 45 features for the multiclass task and 30 for the binary classification task.
- Training set is augmented using five distinct oversampling and synthetic sample generation techniques before classifier fitting: SMOTE, GAN, ADASYN, MIX UP, and RO**.**

IV. LEARNING MODEL SELECTION AND EVALUATION STRATEGY

XGBoost was selected as the classification model due to its high predictive performance and efficiency in handling high-dimensional tabular network data. Following data preparation and augmentation, a separate XGBoost model was trained for each augmentation technique as well as for the original non-augmented dataset. As illustrated in Figure 1, model performance was evaluated under two experimental settings: intra-dataset validation using the CIC-IDS2017 test set and cross-dataset validation using the unseen CSE-CIC-IDS2018 dataset to assess generalization. Performance was measured using Precision, Recall, F1-score, and macro-average values, providing a comprehensive evaluation of both overall classification performance and minority attack detection.

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the results obtained for both binary and multiclass classification.

1. Multi-Class Classification Results

Table 1 presents the multiclass classification results obtained without data augmentation. In the intra-dataset evaluation on CIC-IDS2017, the model achieved consistently high recall values, reaching 0.99 for most attack classes and 0.94 for Web Attack, with a macro recall of 0.98. This strong performance indicates that the model effectively learned the attack patterns represented in the training dataset. However, during cross-dataset evaluation on CIC-IDS2018, recall decreased substantially, with relatively strong performance only for Patator (0.85) and DDoS (0.71), suggesting that these attack patterns were more transferable across the two datasets. In contrast, DoS, Web Attack, and Bot achieved recalls of 0.54, 0.27, and 0.01, respectively, which may reflect differences in traffic characteristics and attack distributions between the datasets. The resulting macro recall of 0.48 highlights the model's limited ability to generalize under distribution shifts. These findings are further supported by the confusion matrices presented in Figures 2 and 3.

Table 1. Multi-Class Classification Performance: CIC-IDS2017 vs. CIC-IDS2018 (Before augmentation)

Traffic Type	CIC-IDS2017			CIC-IDS2018		
	Precision	**F1 score**	**Recall**	**Precision**	**F1 score**	**Recall**
DoS	0.99	0.99	0.99	0.95	0.69	0.54
DDoS	1.00	0.99	0.99	0.86	0.78	0.71
DoS Hulk	0.99	0.99	0.99	/	/	/
Bot	1.00	0.99	0.99	0.99	0.03	0.01
Web Attack	0.97	0.96	0.94	0.97	0.42	0.27
Patator	0.99	0.99	0.99	0.87	0.86	0.85
PortScan	0.99	0.99	0.99	/	/	/
Macro avg	**0.98**	**0.98**	**0.98**	**0.93**	**0.56**	**0.48**

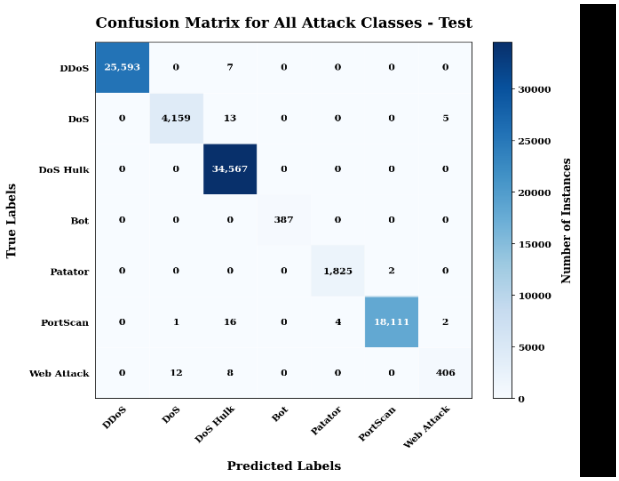


Figure 2. Multiclass confusion matrix of CIC-IDS2017 before augmentation.

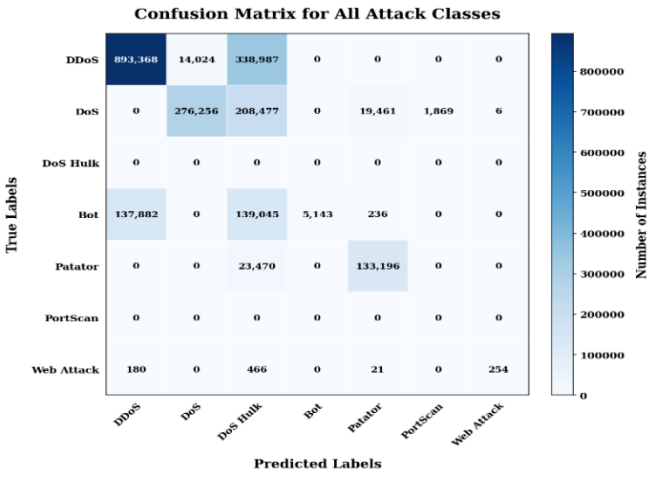


Figure 3. Multiclass confusion matrix of CIC-IDS2018 before augmentation.

Table 2 presents the multiclass classification results obtained after applying various data augmentation techniques. In the intra-dataset evaluation on CIC-IDS2017, all techniques maintained consistently high recall, with a macro recall of 0.99, indicating that augmentation did not degrade performance on the original data distribution. More importantly, augmentation substantially improved cross-dataset performance on CIC-IDS2018 compared with the non-augmented macro recall of 0.48. SMOTE achieved the best generalization performance, with a macro recall of 0.88, followed by RO at 0.84 and MixUp at 0.69, suggesting that balancing minority classes improved the learning of more transferable attack patterns. SMOTE also achieved strong recall for Bot (0.99), DDoS (0.97), Patator (1.00), and DoS (0.85). In contrast, GAN and ADASYN provided only moderate improvements, with macro recalls of 0.60 and 0.54, respectively, suggesting that the generated samples were less representative of the attack distributions in CIC-IDS2018. Overall, these results show that the effectiveness of augmentation varies considerably across techniques, with SMOTE and RO yielding the greatest improvements in cross-dataset generalization. The confusion matrices in Figures 4 and 5 further support these observations.

Table 2. Impact of Data Augmentation Strategies on Attack Detection Recall (CIC-IDS2017 vs. CSE-CIC-IDS2018)

Attack Type	CIC-IDS2017					CIC-IDS2018				
	SMOT	GAN	ADASYN	Mix Up	RO	SMOT	GAN	ADASYN	Mix Up	RO
DoS	0.99	0.99	0.99	0.99	0.99	0.85	0.72	0.69	0.70	0.68
DDoS	0.99	0.99	0.99	0.99	0.99	0.97	0.35	0.37	0.75	0.95
DoS	0.99	0.99	0.99	0.99	0.99	/	/	/	/	/
Bot	1.00	0.99	1.00	1.00	1.00	0.99	0.50	0.50	0.50	0.99
Web	0.99	0.99	0.99	0.99	1.00	0.59	0.45	0.12	0.75	0.60
Patator	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00	0.74	1.00
PortSca	0.99	0.99	0.99	0.99	0.99	/	/	/	/	/
Macro	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.88**	**0.60**	**0.54**	**0.69**	**0.84**

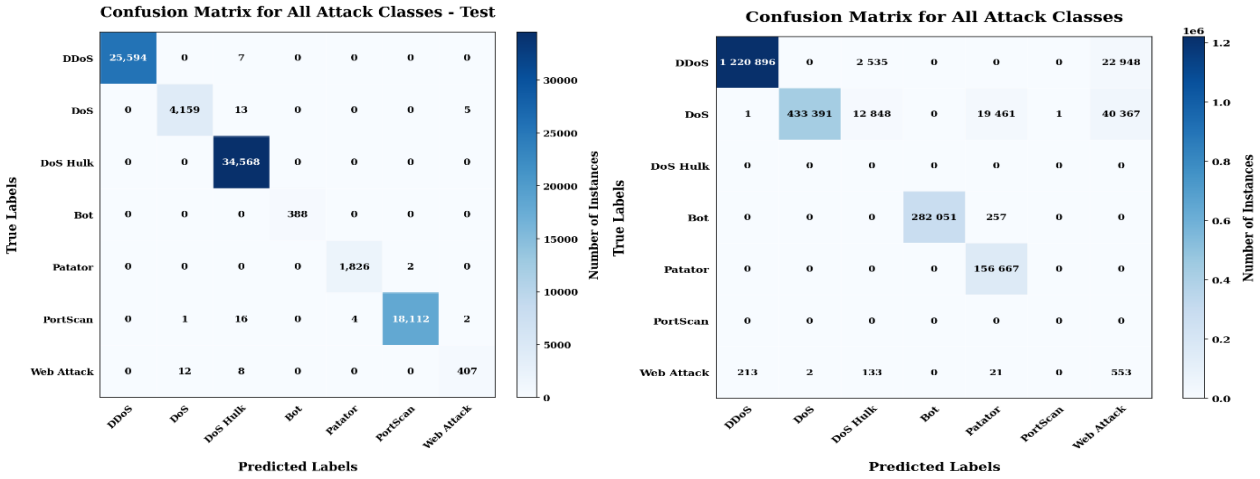


Figure 4. Multiclass confusion matrix of CIC-IDS2017 after augmentation (using SMOTE)

Figure 5. Multiclass confusion matrix of CIC-IDS2018 after augmentation. (using SMOTE)

2. Binary Classification Results

Table 3 presents the binary classification results obtained without data augmentation. In the intra-dataset evaluation on CIC-IDS2017, the model achieved strong performance, with recalls of 0.97 for Benign traffic and 0.95 for Attack traffic, resulting in a macro recall of 0.96. This indicates that the model effectively distinguished malicious from benign traffic when both training and evaluation data originated from the same distribution. However, during cross-dataset evaluation on CIC-IDS2018, the Attack recall decreased substantially to 0.45, indicating that approximately 55% of malicious samples were misclassified as benign. This reduction suggests that the attack patterns learned from CIC-IDS2017 were not fully transferable to the different traffic distribution of CIC-IDS2018. In contrast, Benign recall remained high at 0.99, showing that normal traffic characteristics were more consistently recognized across the two datasets. Consequently, the macro recall decreased to 0.72, highlighting the model's limited ability to generalize attack detection across heterogeneous network environments. These observations are further reflected in the confusion matrix presented in the figure. 6.

Table 3. Comparative Evaluation of Binary Classification Performances: CIC-IDS2017 vs. CIC-IDS2018. (Before augmentation).

Traffic Type	**CIC-IDS2017**			**CIC-IDS2018**		
	Precision	**F1 score**	**Recall**	**Precision**	**F1 score**	**Recall**
Benign	0.99	0.98	0.97	0.81	0.89	0.99
Attack	1.00	0.97	0.95	0.98	0.61	0.45
Macro avg	**0.99**	**0.97**	**0.96**	**0.90**	**0.75**	**0.72**

Table 4 shows that all binary classification models maintained strong intra-dataset Attack recall on the CIC-IDS2017 dataset, with values ranging from 0.97 to 0.99. In contrast, performance decreased substantially on the unseen CSE-CIC-IDS2018 dataset. Nevertheless, most augmentation techniques improved the cross-dataset Attack recall compared with the non-augmented baseline of 0.45. SMOTE achieved the highest recall of 0.78, followed by random oversampling (RO) at 0.70, GAN at 0.60, ADASYN at 0.56, and MixUp at 0.45. The superior cross-dataset performance achieved with SMOTE is further illustrated by the confusion matrix shown in the Figure. 7.

Table 4. Impact of Data Augmentation Strategies on Binary Classification Recall (CIC-IDS2017 vs. CIC-IDS2018)

Traffic Type	CIC-IDS2017					CIC-IDS2018				
	SMOTE	GAN	ADASYN	Mix Up	RO	SMOTE	GAN	ADASYN	Mix Up	RO
Benign	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.95	0.94	0.97
Attack	0.98	0.97	0.99	0.98	0.97	0.78	0.60	0.56	0.45	0.70
Macro Recall	**0.98**	**0.98**	**0.99**	**0.98**	**0.98**	**0.87**	**0.79**	**0.75**	**0.69**	**0.83**

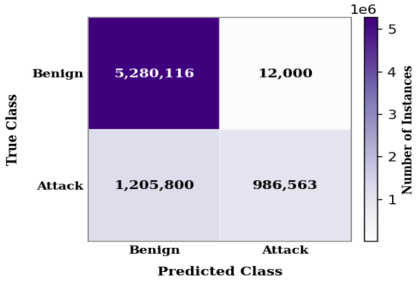


Figure 6. Binary classification confusion matrix of CIC-IDS2018 (Before Augmentation)

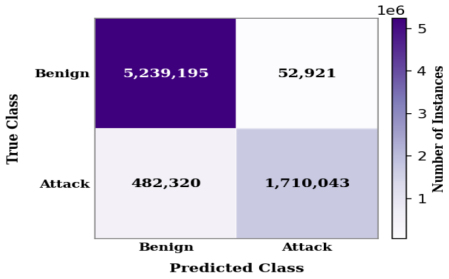


Figure 7. Binary classification confusion matrix of CIC-IDS2018 (using SMOTE)

The findings demonstrate a substantial difference between intra-dataset and cross-dataset performance. While all models achieved high Attack recall on the CIC-IDS2017 dataset, their effectiveness declined markedly when tested on the unseen CSE-CIC-IDS2018 dataset, indicating limited generalization across distinct network traffic distributions. Most augmentation techniques reduced this performance degradation, confirming their potential to improve cross-dataset robustness. SMOTE achieved the best results, followed by RO, suggesting that increasing minority-class representation helped the model learn more transferable attack patterns. In contrast, MixUp provided no improvement over the non-augmented baseline, indicating that interpolation-based augmentation may not preserve the complex characteristics of network attack traffic. Overall, the findings demonstrate that the effectiveness of data augmentation strongly depends on the selected technique.

VI. CONCLUSION

This study systematically investigated the effect of data augmentation on the intra-dataset and cross-dataset performance of ML-based IDS using CIC-IDS2017 for training and CSE-CIC-IDS2018 for external evaluation. The results showed that models achieving very high performance under intra-dataset evaluation can experience substantial degradation when applied to an unseen dataset, confirming that cross-dataset generalization remains a critical challenge for ML-based IDSs. This finding also highlights the limitations of relying solely on intra-dataset evaluation when assessing model robustness. The experimental results further demonstrated that data augmentation can substantially improve cross-dataset detection performance while preserving strong intra-dataset results. Among the five evaluated techniques, SMOTE and random oversampling consistently achieved the greatest improvements, whereas GAN-based augmentation, ADASYN, and MixUp showed more variable effectiveness across evaluation settings and attack classes. These differences indicate that the benefits of augmentation depend strongly on the technique used and the characteristics of the underlying attack

distributions. Moreover, the remaining performance gap across several attack classes suggests that class imbalance alone does not fully explain the generalization problem; distributional shifts and differences in traffic and attack characteristics between datasets also play important roles. Overall, the findings demonstrate that appropriately selected data augmentation techniques can enhance the robustness and transferability of ML-based IDSs across heterogeneous network environments. Still, augmentation alone is insufficient to overcome cross-dataset distribution shifts fully. The results, therefore, provide useful guidance for researchers and IDS developers when selecting augmentation strategies for models intended to operate beyond a single benchmark dataset. Future work will explore hybrid approaches that combine advanced data augmentation with domain adaptation, representation learning, distribution alignment, and multi-dataset training to further improve intrusion detection reliability under previously unseen network conditions.

REFERENCES

[1] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, “Network intrusion detection system: A systematic study of machine learning and deep learning approaches,” Trans. Emerg. Telecommun. Technol., vol. 32, no. 1, Art. no. e4150, Jan. 2021.

[2] V. Shanmugam, R. Razavi-Far, and E. Hallaji, “Addressing class imbalance in intrusion detection: A comprehensive evaluation of machine learning approaches,” Electronics, vol. 14, no. 1, Art. no. 69, Jan. 2025,

[3] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 2018, pp. 108-116.

[4] S. Bourou, A. El Saer, T.-H. Velivassaki, A. Voulkidis, and T. Zahariadis, "A review of tabular data synthesis using GANs on an IDS dataset," Information, vol. 12, no. 9, Art. no. 375, 2021.

[5] M. Al-Ajlan and M. Ykhlef, "A review of generative adversarial networks for intrusion detection systems: Advances, challenges, and future directions," Computers, Materials & Continua, vol. 81, no. 2, pp. 2053-2076, 2024.

[6] Z. Zhao, A. Kunar, R. Birke, H. van der Scheer, and L. Y. Chen, "CTAB-GAN+: Enhancing tabular data synthesis," Frontiers in Big Data, vol. 6, Art. no. 1296508, 2024.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[8] X. Zhao, K. W. Fok, and V. L. L. Thing, "Enhancing network intrusion detection performance using generative adversarial networks," Computers & Security, vol. 145, Art. no. 104005, 2024.

[9] S. Layeghy, M. Baktashmotlagh, and M. Portmann, “DI-NIDS: Domain invariant network intrusion detection system,” Knowledge-Based Systems, vol. 273, Art. no. 110626, Aug. 2023.

[10] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," IEEE Trans. Knowl. Data Eng., vol. 35, no. 8, pp. 8052-8072, Aug. 2023.

[11] J. Mai, C. Gao, and J. Bao, "Domain generalization through data augmentation: A survey of methods, applications, and challenges," Mathematics, vol. 13, no. 5, Art. no. 824, 2025.

[12] A. H. Ali, M. Charfeddine, B. Ammar, F. Albalwy, and A. Hussain, "Unveiling machine learning strategies and considerations in intrusion detection systems: A comprehensive survey," Frontiers in Computer Science, vol. 6, Art. no. 1387354, 2024.

[13] U. C. Akuthota and L. Bhargava, “The role of machine and deep learning in modern intrusion detection systems: A comprehensive review,” Comput. Electr. Eng., vol. 124, Part A, Art. no. 110318, May 2025.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785-794.

[15] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 2018, pp. 108-116.

[16] M. Cantone, C. Marrocco, and A. Bria, "On the cross-dataset generalization of machine learning for network intrusion detection," arXiv:2402.10974, 2024.

[17] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: Latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, Art. no. 137, May 2024.

[18] M. Al-Ajlan and M. Ykhlef, "A review of generative adversarial networks for intrusion detection systems: Advances, challenges, and future directions," Computers, Materials & Continua, vol. 81, no. 2, pp. 2053-2076, 2024.

[19] A. Demirbaş Paray and M. Aydos, "Federated learning for intrusion detection under class imbalance: A multi-domain ablation study with per-client SMOTE," *Appl. Sci.*, vol. 16, no. 2, Art. no. 801, Jan. 2026.

[20] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN), Hong Kong, China, 2008, pp. 1322-1328.

[21] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019.

[22] W. Tian, Y. Shen, N. Guo, J. Yuan, and Y. Yang, "VAE-WACGAN: An improved data augmentation method based on VAEGAN for intrusion detection," Sensors, vol. 24, no. 18, Art. no. 6035, 2024.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2018.

[24] M. Iwanowski, D. Olszewski, W. Graniszewski, J. Krupski, and F. Pelc, "The choice of training data and the generalizability of machine learning models for network intrusion detection systems," Applied Sciences, vol. 15, no. 8, Art. no. 8466, 2025.

[25] M. Cantone, C. Marrocco, and A. Bria, "Machine learning in network intrusion detection: A cross-dataset generalization study," in Proc. IEEE Int. Conf. (ACCESS), 2024, pp. 14489-14508.

[26] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy (ICISSP), 2018, pp. 108-116.

[27] Canadian Institute for Cybersecurity, University of New Brunswick, "CSE-CIC-IDS2018 on AWS," CSE-CIC-IDS2018 Dataset.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Selective Data Cleaning with Large Language Models for Improving E-Commerce Product Matching Training Data

Oğuz Akif Tüfekçioğlu
R&D Department
Cimri Inf. Tech. Syst. Inc.
Istanbul, Türkiye
akif.tufekcioglu@cimri.com

Özgür Aslan
R&D Department
Cimri Inf. Tech. Syst. Inc.
Istanbul, Türkiye
ozgur.aslan@cimri.com

Tuna Çakar
Computer Engineering
MEF University
Istanbul, Türkiye
cakart@mef.edu.tr

Abstract
E-commerce product-matching models are affected by false-positive matches and category-dependent label inconsistencies. This paper presents a selective data-cleaning method to improve human-operator-labeled training data in a large industrial catalog setting. Instead of sending the full dataset to a large language model, a Turkish BERT-based matching model first identifies high-risk records in which the operator label and the model prediction strongly disagree. Only these records are reviewed by Qwen3-8B under category-specific matching rules. In an experiment with 1,559,716 labeled product pairs and 3456 categories, the BERT pre-filter reduced the review set to 13,373 candidates, reducing the large language model review volume by 99.14 percent. Based on the BERT and large language model agreement, 7,586 records were flagged, and 37,659 records were removed after including related negative examples. Retraining on the cleaned data increased F1 from 88.18 percent to 90.18 percent, accuracy from 95.30 percent to 96.09 percent, and AUC from 98.61 percent to 98.95 percent. An ablation against a BERT-only cleaning rule shows that the large language model review is what preserves recall: BERT-only filtering removes 1.72 times as much training data and reaches 89.48 percent F1, compared with 90.18 percent for the combined rule. The results show that large language models can serve as targeted, rule-guided, and cost-aware quality control layers for large-scale e-commerce data cleaning.

Keywords: e-commerce product matching; data cleaning; large language models; BERT; noisy labels.

I. INTRODUCTION

In product and price comparison platforms, the same product can be listed by different sellers with different titles, descriptions, prices, and images. This makes product matching a central data-quality problem, not just a search or classification task. Wrong matches can show different products as the same item, distort price comparisons, feed recommender systems with wrong signals, and reduce user trust. Therefore, model success depends not only on architecture, but also on accurate, consistent, and category-aware training data. In the industrial setting examined in this study, product and offer records are updated frequently. Product titles are often short, incomplete, marketing-oriented, or seller-specific. Different stores may use different names for the same product, while different products may look very similar because they share the same brand, model family, capacity, package content, or color. Past operator decisions are valuable training data, but they may also contain inconsistencies because category rules change and borderline cases are interpreted differently. In the current machine-learning-based matching infrastructure, model training primarily relies on manually labeled data from past operator decisions. If this data contains false positive matches, the model can learn incorrect patterns, especially for product pairs that are textually similar but refer to different product identities [1-5]. Large language models can interpret semantic similarity, follow category rules, and produce explainable decisions. However, reviewing millions of product-matching records using a large language model is impractical due to processing time and hardware costs. This study, therefore, uses the large language model as a selective quality-control layer for records that the existing matching model flags as suspicious. The main hypothesis is that high-confidence conflicts between the operator label and the model output can indicate error regions in the training data. A

Turkish BERT-based model first scans the large dataset; Qwen3-8B then rechecks only the risky candidates under category-specific rules. The paper makes four contributions. First, it presents a label-model disagreement flow to identify false-positive candidates in product-matching data. Second, it uses the large language model as a second-review layer for high-risk candidates, not as a full-data relabeling tool. Third, it applies cleaning with group consistency by removing related negative examples when needed. Fourth, it validates an integrated, cost-controlled industrial architecture that combines BERT-based matching, noisy-label checking, and large-language-model verification. Thus, the main contribution lies in scalable system design, domain rule transfer, and measurable impact on data quality, rather than in introducing a new base learning algorithm.

II. RELATED WORK

Product matching, also known as entity matching, decides whether records from different sources refer to the same real-world object. Classical methods rely on character similarity, field-level thresholds, and rules, whereas deep learning methods learn more flexible representations for short and noisy texts [2], [6]. In e-commerce, product titles are often not standardized and may mix identity information with missing attributes or marketing phrases. Pretrained language models help capture relations between fields such as brand, model, series, variant, and product family [1], [3]. BERT-based models can encode two product titles together and evaluate more than word overlap. However, this advantage depends on reliable labels. If the training data contains false positives, the model may associate semantic similarity with an incorrect matching signal. The noisy-label literature shows that wrong labels can have lasting effects on deep models [4], [5], [7]. Confident learning uses conflicts between model predictions and given labels to estimate possible label errors [4]. Weak supervision and data programming methods, such as Snorkel, combine expert rules, labeling functions, and data sources to create training data more quickly [8], [9]. The present study connects such label-error checking to a second review by a large language model within a high-volume product-matching pipeline. Large language models have also been studied as tools for data preprocessing, cleaning, annotation, and synthesis [10-15]. They can combine language knowledge with task instructions, but running them on all records is costly in high-volume e-commerce settings. The proposed approach is selective: the large language model is applied only to a small candidate subset first marked as risky by a faster base model. In this context, the study combines ideas from confident learning, weak supervision, and large language model-based data preparation in a category-sensitive product matching problem. BERT is used both as a classifier and as a scanning component that produces risky candidates; the large language model serves as a limited-scope, rule-based verification layer. The difference from full LLM relabeling is operational: the expensive semantic reviewer is called only after a fast model has narrowed the error region.

III. PROBLEM DEFINITION

Each sample represents a binary matching decision between an offer product title and a main product title. A positive label means that the titles refer to the same product identity; a negative label means that they do not. Some positive labels can be wrong, for example, when two products share the same brand but differ in capacity, weight, shade, package contents, or main-product vs. accessory status. This study focuses on false-positive matches because they can teach the model to treat different products as the same item. A false positive can also affect negative examples generated around the same offer. Therefore, cleaning is designed as a product-group-level intervention rather than a simple row-level correction. The research question has two parts. First, can operator-based false positives be detected at scale by combining high-confidence BERT disagreements with category-rule-guided large language model review? Second, does this cleaning improve the product matching model when it is retrained under the same evaluation conditions? The problem is formulated as a binary classification problem with label checking. For each record, the title pair is xi, the operator label is yi, and the BERT non-match probability is pi. A record becomes a cleaning candidate only if yi = 1 and pi >= 0.90. This

rule focuses the method on the subset with the highest false-positive risk and keeps the large-language-model review volume low.

IV. DATASET AND DOMAIN CONSTRAINTS

The experiment used an industrial dataset comprising 1,559,716 operator-labeled product-matching samples and 3,456 categories. The samples cover 314,869 distinct offers. Each offer contributes exactly one positive pair and about 4 linked negative pairs derived from variant product titles, yielding a mean group size of 4.95 rows. Each record contains an offer product title and a main product title. Titles are mainly Turkish but may include brand names, model codes, foreign technical terms, units, package counts, and seller-specific abbreviations, which makes simple text similarity insufficient. Category structure is a main source of uncertainty. In electronics, model code or capacity may determine the decision; in cosmetics, shade, volume, and package content may be critical; in fashion, color and size may be variants in some cases and identity separators in others. For this reason, the large language model prompts include category-specific rules rather than asking only a general, same-product question. Positive matches and related negative examples appear together. For one offer, the correct main product can be accompanied by similar but non-matching products. When a positive match is found to be incorrect, related negatives may also lose their meaning, so the cleaning process considers linked records. During preprocessing, product titles were standardized while preserving meaningful fields. Extra spaces, repeated separators, promotional expressions, and differences in formal writing were reduced. Still, model code, capacity, series, package count, volume, and measurement units were kept because these details may define product identity. To ensure evaluation reliability, cleaning was applied only to the training data, and the evaluation split was held fixed. Thus, observed changes could be linked to a more consistent training signal rather than to a changed test set. The dataset is industrial and cannot be shared publicly; therefore, the paper reports the observable scale, category coverage, cleaning counts, and performance deltas, while treating detailed category distributions and some internal annotation details as reproducibility limitations. One structural property of the splits should be noted. Offers are cleanly partitioned: none of the 314,869 offers appears in more than one split. Products are not: 32,356 of 84,621 distinct products appear in more than one split, because different offers matched to the same product can fall on different sides of the partition. Re-splitting the data on product identity would remove this overlap but would also invalidate the comparison with the existing production model, so it is left as future work.

Table 1. Example matching rules that differ by category.

Category example	Critical fields	Effect on the decision
Electronics	Model code, capacity, series	Strong veto if different
Cosmetics	Shade, volume, package content	Variant and amount check
Pet products	Animal type, weight, count, age group	Correct type and package required
Fashion	Size, color, gender, and collection	Category-dependent flexibility
Accessories	Main product vs. compatible part	Wrong product type veto

V. METHOD

The selective data cleaning method has three layers. First, a Turkish BERT-based product-matching model is trained on the initial operator-labeled dataset and then re-run on the same data. Records labeled as positive by the operator but predicted as non-matches with high confidence are accepted as risky candidates.

A 90 percent non-match probability threshold was used for candidate selection. This conservative point sends only strong BERT-operator disagreements for large-language-model review, thereby limiting costs and reducing the risk of deleting true matches. Keeping the threshold high narrows the candidate pool. Lower thresholds may find more errors but increase review volume and deletion risk; higher thresholds reduce cost but leave more false positives. This study keeps the threshold fixed at 0.90. Two observations support the choice. First, the large language model rejection rate increases monotonically with BERT confidence, from 49.5 percent in the 0.90-0.925 band to 69.5 percent above 0.99, indicating that the pre-filter score carries meaningful information rather than serving as an arbitrary cutoff. Second, raising the threshold to 0.95 reduces the candidate pool from 13,373 to 8,848 while raising the rejection rate only from 57.7 to 60.9 percent, so the extra strictness mostly discards true errors. A full sensitivity analysis, with a separate retraining run for each threshold, is left for future work.

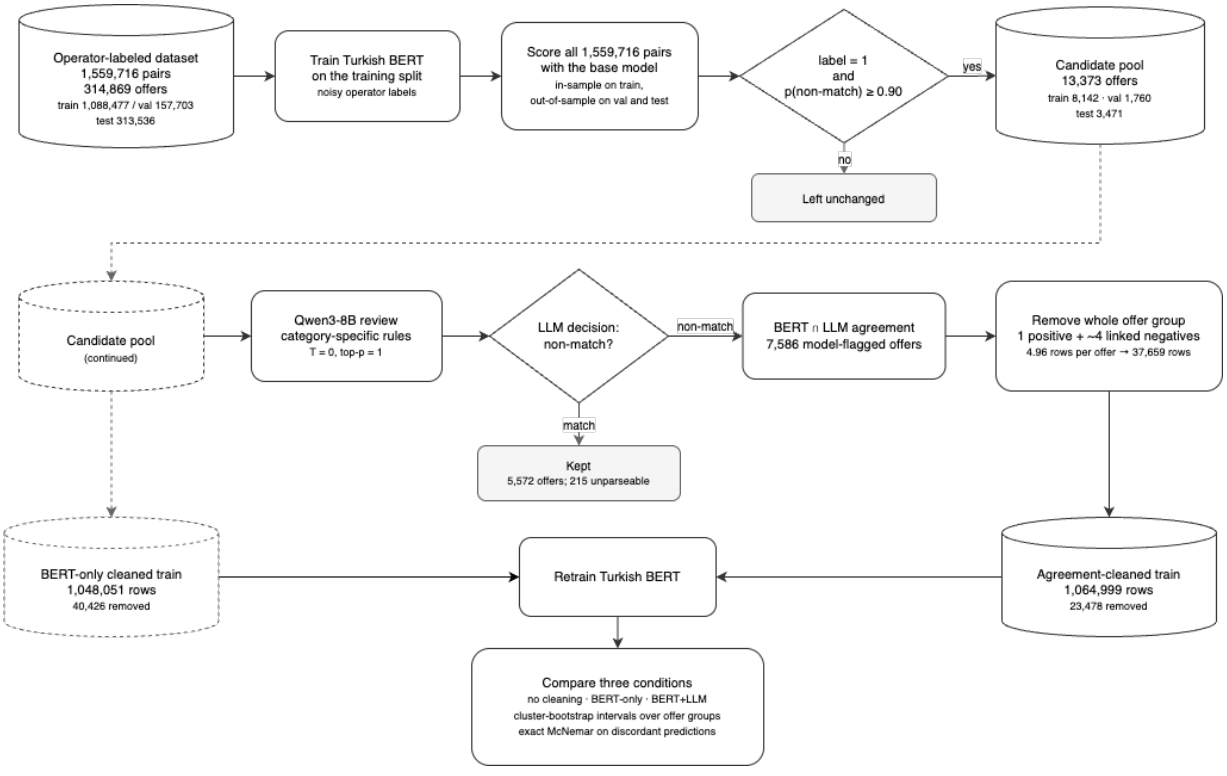


Figure 1. The selective data cleaning pipeline was executed. A single Turkish BERT model is trained on the noisy training split and then scored on the whole dataset. High-confidence disagreements with the operator label become candidates, which are reviewed by Qwen3-8B under category-specific rules. Removal occurs at the offer group level. All conditions are then retrained and compared on the same evaluation split. The dashed path is the BERT-only ablation, which skips the review step.

Second, Qwen3-8B is guided by category-specific rules. The prompt defines which fields break a match, which differences can be accepted as variants, and which cases must be marked as non-matches. This prevents the model from relying only on surface text similarity. Third, BERT and large-language-model outputs are evaluated

together. A record is removed only if both layers mark it as a non-match. BERT serves as a fast suspicion generator, while the large language model serves as a slower but more detailed, rule-based reviewer. When a cleaning decision is made, group consistency is preserved. The false positives and their linked negatives are removed together so that the model does not learn inconsistent signals around the same offer.

VI. LARGE LANGUAGE MODEL REVIEW

The large language model review is not a final production decision module. It is a semi-automatic offline review layer for improving training data. This keeps production matching fast while limiting the more expensive large-language-model step to the most useful subset of candidates. The prompt design has two steps. First, the category context is identified. Second, category rules are used to decide whether the titles describe the same product, an acceptable variant, or different products. The decision checks the brand, model, technical fields, measurement units, package content, category-specific meanings of color or size, and main product/accessory differences. The prompt asks for three outputs: a match or non-match decision, a short list of critical fields, and a veto reason if a critical difference invalidates the match. In compact form, the prompt contains the offer title, the main product title, the detected category, category-specific critical fields, allowed variant differences, veto rules, and the requested structured output. This creates auditable decision traces instead of long free-form explanations. The review explicitly avoids depending only on textual closeness. Sellers may list the same product family under different campaign words or with missing fields. At the same time, small differences, such as 128 GB versus 256 GB or single versus multi-pack content, can change the product identity. Therefore, category-specific tolerance and veto rules are used, and uncertain cases are cleaned only when BERT and large language models agree. The large language model output is treated as structured decision logic. The expected answer is the decision, the critical conflict, and the related category rule. This helps operators verify the result and refine category prompts as needed. This setup adds category knowledge to the data cleaning loop. BERT may treat similar brand-and-series phrases as positive evidence. At the same time, the large language model can reject the match when capacity, model code, or package content violates the category rule.

VII. EXPERIMENTAL SETUP

The baseline is a Turkish BERT-based matching model that takes offer-main product title pairs as binary classification input. It was trained on 1,088,477 operator-labeled samples and then applied back to the data to find positive labels it strongly disagreed with. In the BERT model, the offer title and the main product title are processed as a pair of inputs separated by special tokens. The classification head produces a match or non-match decision from the pair representation. The model is fine-tuned on the task data rather than used solely as a frozen feature extractor, and evaluation is performed on a fixed split separate from the cleaning process. The same architecture and evaluation split are used before and after cleaning. All runs fine-tune dbmdz/bert-base-turkish-uncased for two epochs with a per-device batch size of 64, gradient accumulation of 4, a learning rate of 5e-5, 500 warmup steps, a maximum sequence length of 512, and mixed precision. The checkpoint with the best validation precision is kept. The reviewer model is Qwen3-8B in GGUF format served through llama.cpp, decoded with temperature 0.0, top-p 1.0, and a 1024-token limit. No random seed was fixed in any run, so exact reproduction of a single run is not possible, and run-to-run variance was not measured. Multiple seeds and cross-validation are not included in this version.

Table 2. Candidate filtering and cleaned records in the data cleaning flow.

Measure	Value	Explanation
Labeled samples	1,559,716	Operator-labeled product pairs

Measure	Value	Explanation
Positive labels	314,869	Only positive labels can become candidates
Number of categories	3456	Broad category coverage
BERT risky candidate pool	13,373	0.86 percent of the dataset, 4.25 percent of positive labels
Records flagged by the BERT and LLM agreement	7,586	BERT and LLM agreement (56.7 percent of the candidate pool)
Records removed with related negatives	37,659	Records removed while preserving group consistency
LLM review volume reduction	99.14 percent	Selective review instead of full-data review

The BERT scan reached about 6577 sequences per second on RTX 5090 hardware, so a full pass over the 1,559,716-sample dataset takes about four minutes. With a 90 percent non-match threshold, 13,373 records were selected, so the large language model reviewed only about 0.86 percent of the dataset, or 4.25 percent of the positive labels. Because the pre-filter itself is inexpensive, almost all of the method's processing cost is due to the large language model review time, which is controlled by the selectivity ratio. Qwen3-8B was used for the large-language-model review. With llama.cpp on a single GPU, the observed speed was about 45 tokens per second. Since the model reviewed only BERT-selected candidates, the candidate pool size became the main driver of processing cost. The evaluation used three measurement groups: reduction in large language model review volume, model performance after retraining under each cleaning condition, and practical processing time. Processing time is reported from measured throughput rather than estimated. Three conditions are compared quantitatively in Table 3: no cleaning, cleaning using the BERT pre-filter alone, and cleaning using BERT and large-language-model agreement. A full large language model pass over every record was not run because, at the observed review speed, it would require several weeks of processing for this dataset.

VIII. RESULTS

Of the 13,373 candidates sent for review, the large language model rejected 7,586, accepted 5,572, and returned 215 responses that could not be parsed. The two layers, therefore, agree on 56.7 percent of the candidate pool. This is an agreement rate between two automated layers, not a detection rate. No human-validated reference set was built for this study, so the number of mislabeled positives outside the flagged region is unknown, and no claim about recall can be made. For the same reason, the 7,586 records are described throughout this paper as model-flagged rather than as confirmed errors.

As summarized in Table 2, from 1,559,716 samples, 13,373 records were identified as risky by BERT.

After a large language model review, 7,586 records were flagged by BERT and the large language model agreement. With related negatives included, 37,659 records were removed, reducing the review volume by 99.14 percent compared with the full-data review.

Table 3. Performance of the Turkish BERT matching model under three cleaning conditions. All conditions are scored on the same evaluation split of 304,101 pairs from 61,377 offers. Removal counts are training rows.

Condition	Train rows	Removed	Precision	Recall	F1	Accuracy	AUC
No cleaning	1,088,477	0	89.45	86.94	88.18	95.30	98.61
BERT-only, threshold 0.90	1,048,051	40,426	91.69	87.37	89.48	95.85	98.74
BERT and LLM agreement	1,064,999	23,478	91.47	88.93	90.18	96.09	98.95

The Turkish BERT model was retrained under three cleaning conditions and compared on the same evaluation set. Under the BERT and large language model agreement, precision increased from 89.45 percent to 91.47 percent, recall from 86.94 percent to 88.93 percent, F1 from 88.18 percent to 90.18 percent, accuracy from 95.30 percent to 96.09 percent, and AUC from 98.61 percent to 98.95 percent. The joint increase in precision and recall suggests that cleaning did not simply make the model more conservative. Removing false positives and related negatives helped the model reduce wrong matches while keeping correct matches more consistent. The 2.27 percent relative F1 increase indicates a measurable improvement. The BERT-only condition isolates the contribution of the large language model layer. Removing every high-confidence BERT disagreement without a second review yields the highest precision among the three conditions: 91.69 percent, slightly above the combined rule. It recovers almost none of the recall, however: 87.37 percent against 88.93 percent, and it ends 0.70 points behind in F1. The reason is visible in the removal counts. BERT-only deletes 40,426 training rows against 23,478, and the additional 16,948 rows cover 3,415 offers, 42 percent of everything the pre-filter flagged, that the large language model examined and judged to be genuine matches. Those offers carry the borderline-but-correct patterns the model needs to maintain recall. Skipping the review, therefore, removes 1.72 times as much training data and produces a more conservative model that misses more true matches. The AUC gain is smaller because the starting AUC was already high. For the BERT-only condition, per-record predictions were retained, so the comparison against the uncleaned baseline can be tested directly. Confidence intervals were computed using a cluster bootstrap with 2000 replicates that resamples entire offer groups rather than individual rows, because the roughly 5 rows belonging to one offer are not independent. All five metric gains exclude zero: precision +2.24 points, 95 percent interval [+2.06, +2.43]; recall +0.42 [+0.22, +0.63]; F1 +1.30 [+1.16, +1.44]; accuracy +0.56 [+0.50, +0.61]; and AUC +0.13 [+0.10, +0.16]. An exact McNemar test over discordant predictions yields 3,167 records correct only under the baseline, compared with 4,861 correct only under BERT-only cleaning, with a two-sided p-value below 1e-70. Per-record predictions were not available for the agreement condition in this analysis, so the same intervals could not be computed for it. Because no random seed was fixed, run-to-run variance was not measured, and multiple-seed replication remains necessary to assess the stability of the reported gains.

IX. DISCUSSION

The results show that a large language model can be useful as a targeted review layer to improve training data, but not as a direct production decision model. BERT provides speed and scalability, while the large language model provides semantic reasoning and rule-following ability. A key advantage is that the cost of the large language model does not grow directly with the full data volume. Only 13,373 of 1,559,716 records were reviewed by the large language model. This is more sustainable than full-dataset relabeling and shows that large language models can create value by giving a second opinion only where existing models struggle. Operationally, the selective flow creates about a 116.6-fold reduction in review volume compared with running the large language model on all data. This reduces compute time, output checking, decision logging, error analysis, and possible human approval effort. Thus, even a limited gain in accuracy can be meaningful when

combined with higher precision, higher recall, and lower review volume. The group-consistency cleaning strategy is also important in practice. In product matching data, positive and negative samples often gain meaning together around the same offer. If a false-positive match is removed but related negatives are retained, the model may still receive inconsistent signals in the same context. For this reason, the proposed method treats the cleaning decision at the level of the training example group, not only at the row level. This leads to a more consistent data improvement process. The approach also has limitations. First, the quality of large language model decisions depends on how complete, clear, and up to date the category rules are. If the rules are missing or unclear, the model may rely too much on surface similarity. Second, the study primarily focuses on cleaning up false-positive matches. Finding false negatives may require different sampling strategies. A third limitation concerns candidate generation. The pre-filter is trained on the training split and then scored back on it, so its predictions on training records are in-sample and prone to memorization. The size of this effect can be measured directly because the validation and test splits were never used for training. At the 0.90 threshold, the model flags 3.70 percent of training positives, 5.53 percent of validation positives, and 5.49 percent of test positives. Validation and test agree within 0.04 points, indicating memorization rather than a distributional difference, and implying that in-sample scoring suppresses the flag rate by a factor of 1.48. The direction of this bias is conservative: it under-detects rather than over-detects, so the candidate pool used here is smaller than that produced by an out-of-fold pipeline. It is also worth noting that 5,231 of the 13,373 candidates, 39.1 percent, come from the validation and test splits and were therefore already scored out-of-sample. Grouped out-of-fold candidate generation, with folds formed by product identity, is the natural next step and would be expected to surface roughly 12,100 training candidates rather than 8,142. There is also a risk of deleting correct examples by mistake, especially in unclear product variants. For example, if capacity or package amount is missing from the title in the same product family, the model may interpret two records as different products. Similarly, the main product versus accessory distinction, the category-dependent meaning of color and size, or missing product codes in seller titles may lead to wrong cleaning decisions. Therefore, the proposed system uses an automatic but conservative agreement rule. Borderline cases should be sent back to operators or experts and used to improve the rule pool. In addition, the method does not directly target false negative matches. In future work, it can be extended into a two-sided cleaning strategy. For example, records that the operator labeled as non-match but that BERT predicts as match with very high confidence could be sent to a separate active learning queue. However, this extension requires more careful human validation than false-positive cleaning does, because adding a record as a positive match may create incorrect merges in the catalog. The most important operational output of the method is that data quality improvement can be linked to measurable improvements in model performance. A data cleaning study should not be evaluated solely by the number of records removed. It should also show how the cleaned data affects model learning. In this study, retraining under the same evaluation conditions made the effect of data cleaning on model decision quality directly visible. For sustainable use in production-like settings, the decision flow should not depend only on model scores. In the proposed architecture, each cleaning decision can be stored with the source record, BERT score, large language model decision, category rule, removed related negatives, and timestamp. This audit trail can help recover incorrectly cleaned examples, improve category rule sets, and analyze potential changes in data distribution over time. From an ablation and comparison perspective, each component of the proposed architecture has a different role. Cleaning with BERT alone is fast and inexpensive, and Table 3 shows that it already improves the model over no cleaning. It over-deletes, however: it discards 42 percent more offers than the combined rule, including many that a rule-guided review recognizes as correct, and the resulting model gives up 1.56 points of recall and 0.70 points of F1. Reviewing the full dataset solely with the large language model may yield strong semantic decisions, but at the observed review speed, it would require several weeks of processing. The BERT and LLM agreement strategy, therefore, combines the speed of BERT with the decision quality of the large language model, and the ablation indicates that the second review earns its cost by preventing the removal of informative true positives. The 90 percent threshold was selected as a conservative operating point to reduce the risk of deleting correct examples by mistake. Bootstrap confidence intervals and a significance test are reported for the BERT-only comparison, and candidate volumes are reported

for thresholds ranging from 0.80 to 0.95. Threshold sensitivity, separate retraining runs per threshold, multiple-seed experiments, and cross-validation are not reported in this version. In terms of failure analysis, missing capacity information, unclear variants, package content, main product versus accessory distinction, and missing product codes can increase the risk of deleting a correct example. For such borderline cases, human approval, rule updates, and active learning queues are recommended.

X. CONCLUSION

This study presented a selective data-cleaning approach that combines BERT pre-filtering and large-language-model review to clean operator-based false positives in e-commerce product-matching training data. Experiments with 1,559,716 samples and 3456 categories showed that the large language model review volume could be reduced by 99.14 percent. The agreement between BERT and the large language model flagged 7,586 records, and 37,659 records were removed when related negatives were included. After retraining the Turkish BERT model with the cleaned data, all main metrics improved: precision, recall, F1, accuracy, and AUC. The increase in F1 from 88.18 percent to 90.18 percent shows that the proposed method does more than make the data more consistent. An ablation against a BERT-only cleaning rule shows that the large language model review is the component responsible for preserving recall. Filtering on the BERT score alone removes 1.72 times as much training data and reaches 89.48 percent F1, so the second review adds 0.70 points of F1 while deleting substantially less data. Cleaning, therefore, improves both data consistency and the final decision quality of the product matching model. The findings show that large language models can serve as cost-controlled, domain-rule-guided, and measurable quality-control layers for high-volume catalog data. Future work can extend the method with grouped out-of-fold candidate generation, an independently human-annotated validation set, a random-removal control, multimodal product images, active learning-based operator feedback, category-level threshold learning, automatic rule discovery, student model distillation, multiple-seed experiments, and threshold sensitivity analysis. Because the BERT pre-filter already narrows the candidate pool, replacing Qwen3-8B with a smaller Qwen-based student model trained through knowledge distillation may further reduce processing time and hardware cost. The method can also be supported by automatic category-based threshold adjustment, rule updates from operator feedback, and periodic retraining cycles. In such an architecture, the large language model becomes not a one-time data-cleaning tool but a periodic, auditable, and cost-controlled component of catalog quality monitoring.

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proc. NAACL-HLT, pp. 4171-4186, 2019.

[2] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, “Deep learning for entity matching: A design space exploration,” Proc. ACM SIGMOD, pp. 19-34, 2018.

[3] Y. Li, J. Li, Y. Suhara, A. Doan, and W. C. Tan, “Deep entity matching with pre-trained language models,” arXiv:2004.00584, 2020.

[4] C. Northcutt, L. Jiang, and I. Chuang, “Confident learning: Estimating uncertainty in dataset labels,” Journal of Artificial Intelligence Research, vol. 70, pp. 1373-1411, 2021.

[5] H. Song, M. Kim, D. Park, Y. Shin, and J. G. Lee, “Learning from noisy labels with deep neural networks: A survey,” IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 8135-8153, 2022.

[6] N. Barlaug and J. A. Gulla, “Neural networks for entity matching: A survey,” ACM Transactions on Knowledge Discovery from Data, vol. 15, no. 3, pp. 1-37, 2021.

[7] H. Zhang, Y. Zhang, J. Li, J. Liu, and L. Ji, "A survey on learning with noisy labels in natural language processing: How to train models with label noise," Engineering Applications of Artificial Intelligence, vol. 146, 110157, 2025.
[8] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," Proc. VLDB Endowment, vol. 11, no. 3, pp. 269-282, 2017.
[9] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: Creating large training sets, quickly," Advances in Neural Information Processing Systems, vol. 29, 2016.
[10] A. Narayan, I. Chami, L. Orr, S. Arora, and C. Ré, "Can foundation models wrangle your data?" arXiv:2205.09911, 2022.
[11] H. Zhang, Y. Dong, C. Xiao, and M. Oyamada, "Large language models as data preprocessors," arXiv:2308.16361, 2023.
[12] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, et al., "Large language models for data annotation and synthesis: A survey," Proc. EMNLP, pp. 930-957, 2024.
[13] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, et al., "Qwen3 technical report," arXiv:2505.09388, 2025.
[14] H. Zhang, Y. Dong, C. Xiao, and M. Oyamada, "Jellyfish: A large language model for data preprocessing," arXiv:2312.01678, 2023.
[15] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," Proceedings of the National Academy of Sciences, vol. 120, no. 30, e2305016120, 202

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

AI-Powered Smart Sensor for Direct Analysis of Aflatoxins in the Open Air Environment

Fred Paul Mark Jjunju
Department of Computer Science, School of Technology and Computer Engineering. Universal Technology and Management University, Plot 8374, Block 82, Kyadondo-Kungu, Gombe, Nansana, Wakiso, P. O. Box 73307, Kampala, Uganda. fpmjjunju@utamu.ac.ug
ORCID: 0000-0001-6257-434X

Denis Bbosa
Department of Mechanical Engineering, Faculty of Applied Science and Technology.Mbarara University of Science and Technology,
P.O. Box 1410, Mbarara, Uganda
dbbosa@must.ac.ug
ORCID:0000-0002-3547-0868

Abstract
Centralized laboratory-based detection methods for Aflatoxins, such as ELISA and HPLC, are expensive and time-consuming and are also constrained by limited laboratory capacity and sample transportation challenges in developing countries. This bottleneck prevents timely detection and intervention, leaving producers and consumers exposed to the health risks of Aflatoxins. Therefore, the lack of rapid, low-cost, and field-deployable aflatoxin detection tools represents a critical gap. A novel handheld, low-cost sensor for analyzing aflatoxin-related compounds on surfaces in open air, based on an IoT and AI architecture, has been developed and deployed, offering portability for in-field analysis. A sensor array comprising MQ3, MQ5, MQ131, MQ135, MP503, temperature, and humidity sensors, with a micropump to aid airflow across the sensors, was constructed and housed in a 3D-printed box. Custom-designed circuitry within the device was developed, and when powered on, the sensor readings can be sent to the cloud or a local repository via USB or Wi-Fi. Initial results have indicated that the portable AI powered sensor for aflatoxins is a green sensor for field applications, without the use of solvents or gases for the analysis of analyte(s) in neat solution near real-time analysis and quantification using assemble machine learning AI model that combines random forest, SVM and ANN with 85% accuracy and sensitivity obtained for the analysis of Aflatoxins in the open environment with little to no sample preparation in the open air.

Keywords: aflatoxins analysis; AI-powered IoT Sensor; field analysis, real-time analysis; food safety

I. INTRODUCTION

Food contamination is a persistent global challenge with far-reaching implications for human health, food security, and international trade. Among the most concerning contaminants are mycotoxins, particularly aflatoxins, produced by Aspergillus species (e.g., A. flavus and A. parasiticus) and frequently found in staple crops such as maize, peanuts, cereals, and spices [1]. The World Health Organization (WHO) recognizes aflatoxins as among the most poisonous mycotoxins and highlights their genotoxic and carcinogenic effects, including acute poisoning and long-term cancer risks [2]. Globally, about 25% of food crops are estimated to be affected by mycotoxins, with aflatoxins being the most prevalent [3]. In sub-Saharan Africa, environmental conditions, particularly warm and humid climates, foster fungal growth and mycotoxin production, while

constrained post-harvest infrastructure exacerbates the risk [4], [5]. The East African region has witnessed serious outbreaks of acute aflatoxicosis, notably in Kenya (2004) with 317 reported cases and a fatality rate of ~39%, and in Tanzania with 68 cases and a ~30% fatality rate [6].

Aflatoxin contamination in Uganda's food supply poses an urgent threat to public health, food safety, and economic development. A recent survey revealed that 41.8% of groundnut samples and 62.8% of maize products in eastern and northern regions exceeded the European Union limit of 20 ppb for total aflatoxins, a level deemed unsafe for food and feed [7]. Additionally, another study across Masindi and Soroti districts found that 45% of maize samples and 30% of groundnut samples were contaminated with aflatoxins above detection thresholds, highlighting widespread exposure across diverse agro-ecological zones [8].

Chronic exposure to aflatoxins is known to cause severe health issues, including liver cancer, stunting in children, and immune system suppression, amplifying its burden on vulnerable rural communities [9]. Furthermore, contaminated maize and groundnuts are frequently rejected in export markets, limiting trade opportunities and undermining the incomes of smallholder farmers [10]. Currently, Uganda relies on centralized laboratory-based detection methods such as ELISA and HPLC [11]. These methods are not only expensive and time-consuming but also constrained by limited laboratory capacity and challenges in sample transportation. This bottleneck prevents timely detection and intervention, leaving producers and consumers exposed. Therefore, the lack of rapid, low-cost, and field-deployable aflatoxin detection tools represents a critical gap [12]. Addressing this gap is essential for safeguarding public health, enhancing market access, and aligning with national goals such as Uganda Vision 2040, National Development Plan IV (NDP IV), and Sustainable Development Goals (SDGs 2, 3, and 9).

The growing challenge of aflatoxin contamination in food systems, particularly in tropical and subtropical regions, where maize, groundnuts/peanuts, spices, and cereals are highly susceptible [10]. Conventional lab-based analytical techniques, such as HPLC, ELISA, and LC-MS/MS [13], although accurate, are expensive, destructive, and unsuitable for rapid, large-scale monitoring [14]. Recently, artificial intelligence (AI), in its various forms (e.g., supervised, unsupervised, and reinforcement learning), has advanced aflatoxin detection and quantification [3]. By combining state-of-the-art Internet of Things (IoT) and deep learning, and integrating models with blockchain technology, high accuracy and specificity can be achieved in detecting aflatoxins [15]. Here, an AI-powered Internet of Things (IoT) sensor platform for rapid, portable, and integrated AI and machine learning (ML) is developed and applied to detect aflatoxin active compounds in open environments with little or no sample preparation. The developed platform can complement lab-based analytical systems by enabling rapid, non-destructive, and predictive in-field detection. Low-cost detection of food contaminants, focusing on aflatoxins, has been developed and applied to detect active, volatile aflatoxin compounds in the open air. The developed portable sensor is integrated with an AI ensemble model comprising a random forest, an SVM, and an ANN. The developed platform is significant in several ways: (1) Public Health Impact: By enabling real-time detection of aflatoxins at the point of production and storage, the project will contribute to reducing chronic exposure that is linked to liver cancer, child stunting, and immune suppression [9]. Early detection and intervention will directly improve food safety and protect consumers (2). Economic Contribution: Aflatoxin contamination has led to repeated rejections of Ugandan maize and groundnut consignments in international markets, undermining farmers' incomes and national trade competitiveness [9]. A reliable local detection solution will enhance compliance with global standards, open export opportunities, and strengthen the agricultural value chain [16].

II. EXPERIMENTAL SECTION

1. Chemicals and Reagents

All analytical aflatoxin volatile organic chemical standards of 1-Octen-3-ol ("Mushroom Alcohol"), a very common fungal VOC associated with moldy odors and emitted by Aspergillus species. It is used as a biomarker for mold contamination in grains, often with a low odor detection threshold (Inamdar, Arati A., et al.); 3-Methyl-1-butanol and 2-Methyl-1-propanol are alcohols frequently associated with both aflatoxigenic and non-aflatoxigenic A. flavus strains. They are identified as key volatiles in contaminated pistachios and corn [17]; 2-Pentylfuran is often found in the headspace of Aspergillus flavus-contaminated grain and is highly associated with A. flavus infection in samples such as maize/corn [18]. "An overview of fungal volatile organic compounds (VOCs)" Fungal associations; Dimethyl Disulfide is part of the VOC profile found in Aspergillus species-infected samples [19]. All VOC standards were dissolved in HPLC-grade methanol. All working solutions of the samples (1-100 µg/mL) were prepared in methanol to the target concentration by stepwise dilution.

2. Sample Preparation

In all experiments, the sample preparation step was reduced to diluting explosive stock solutions in methanol to the desired concentration. From each solution, 3 µg/mL was applied to cellulose filter paper (Whatman, Maidstone, UK, Grade 1) and analyzed immediately with a portable, AI-powered sensor, as shown in Figure 1.

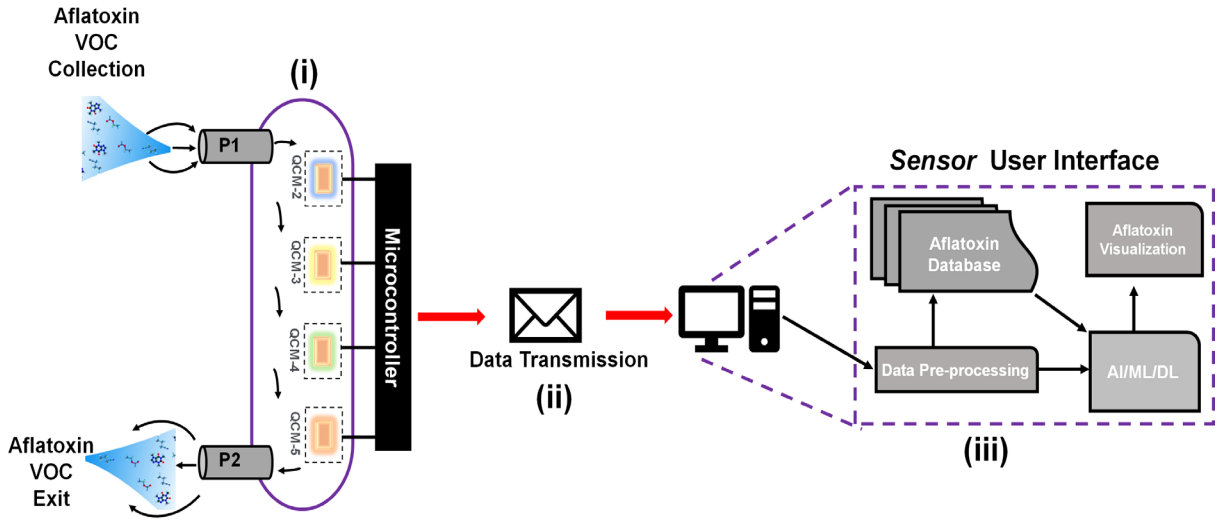


Figure 1. Experimental setup for the direct analysis of Aflatoxins at atmospheric pressure in open air. Schematic of the smart intelligent IoT sensor platform for food contaminants monitoring. The system consists of: (i) array of sensors, (ii) data transmission via Bluetooth, USB to a local PC, Wi-Fi or 4G network to the cloud for visualization and provide informative recommendations for users to take actions to safeguard health and reduce/prevent food contamination, (iii) the collected data can be aggregated securely and in line with data protection best practice and stored in a centralized database for future use to improve strategic understanding and response to poor food quality.

3. AI-powered portable Sensor for Aflatoxin analysis

A custom-designed, handheld, portable, AI-based IoT sensor (Figure 1) was developed. It consists of an array of sensors: MQ3, MQ5, MQ131, MQ135, MP503, temperature, and humidity sensors that were mounted in a 3D printed box (Figure 2 (B)). The sensor array is connected to the custom-designed circuitry module and the microcontroller. A small diaphragm pump (Schwarzer Precision, 12FC35A) with a maximum flow rate of 6 L/min was used to aid in drawing samples from the open environment to the sensing module across the sensor array for detection and to create an air flux within the portable sensor device platform.

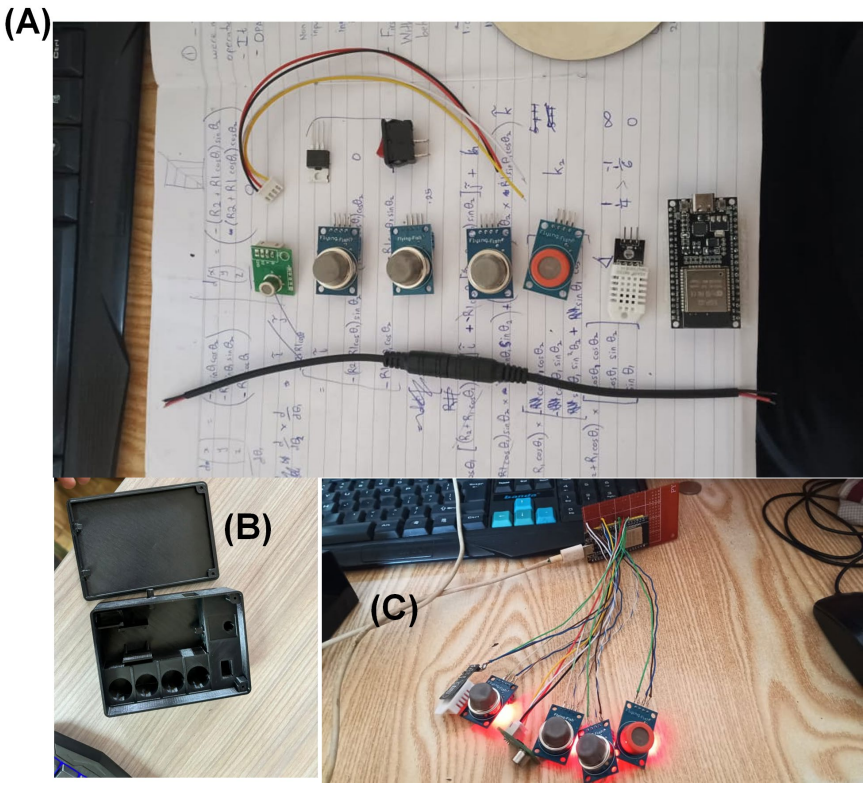


Figure 2. Shows the development of the IoT-based sensor before packaging: (A) list of the sensor components, (B) 3D-printed sensor housing, (C) testing of the individual sensors.

A custom-designed high-voltage control circuit is used to generate a high voltage in the range of 1-5 kV. Figure 2 shows a fully packed portable sensor in a 3D-printed plastic enclosure (MakerBot Replicator 2, Robo Savvy, UK). The circuitry of the portable sensor is powered by a 12 V, 4 Ah lithium Polymer battery with an in-built charging circuit and a power level indicator that indicates when recharging is necessary. From a fully charged state, the battery can power the device for up to 12 hours of continuous use and up to 3-4 days when used intermittently. The high voltage regulated DC-to-DC converter module (EMCO HIGH VOLTAGE, c50/c50N) and the diaphragm pump (Schwarzer Precision, 12FC35A) require a minimum 12V supply to operate correctly; therefore, due to the discharge curve of a battery, it is necessary to include a buck-boost DC-to-DC converter to ensure voltage regulation at the required level at all times. An ARM-based microcontroller board (Teensy 3.1, PJRC, US) handles all internal command communication and receives user input via a laptop. The high-voltage DC-to-DC converter is programmed via a precision voltage reference and a digital potentiometer whose wiper

position is set by commands from the microcontroller. The pump speed is controlled by digitally varying the duty cycle of the PWM (Pulse Width Modulated) signal that drives the diaphragm pump's MOSFET. Figure. Figure 3 shows the entire portable AI-powered sensor housed in a 3D-printed plastic enclosure (MakerBot Replicator 2, Robo Savvy, UK).

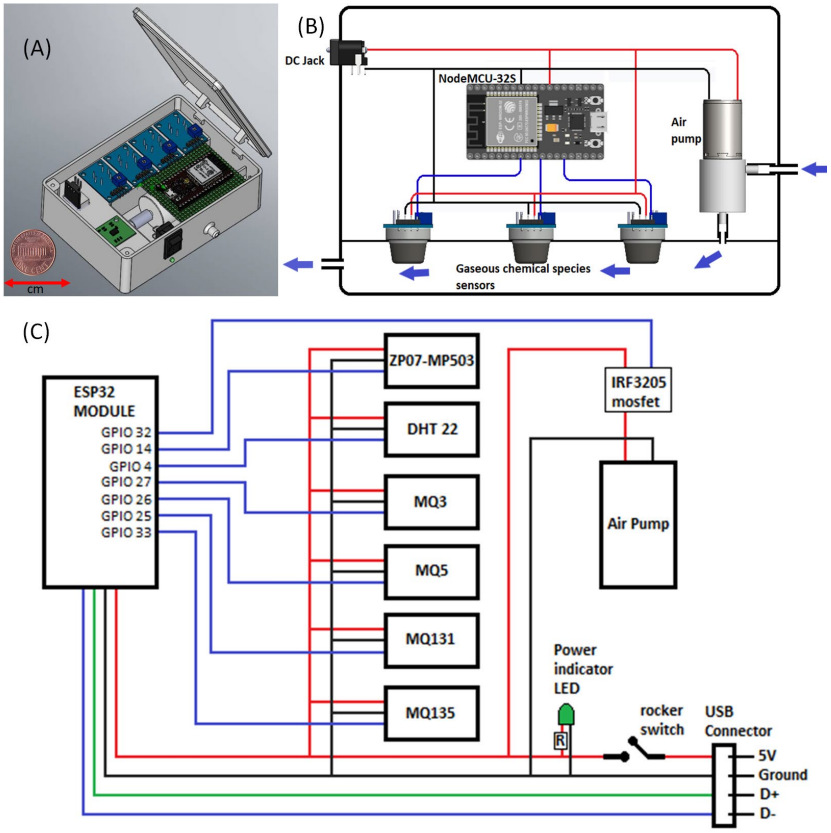


Figure 3. (A) 3D diagram of the packaged Smart Aflatoxin sensor device, (B) shows the layout of the major components of the developed sensor platform for aflatoxin analysis, (C) the electrical Circuit diagram for the developed sensor.

The device operates by sucking in the gases/air to be tested using a small air pump and blowing it over six sensor modules in a channel. It is made up of the following components; An ESP32 microcontroller which serves as the brains of the device, a small 5v DC air pump, an IRF3205 MOSFET that is used by the microcontroller to switch the air pump on or off, a ZP07-MP503 sensor, a DHT22 temperature and humidity sensor, an MQ3 sensor, an MQ5 sensor, an MQ131 ozone sensor, and an MQ135 sensor. The type of sensors is chosen based on their ability to detect a specific chemical class. The device has a rocker switch to turn it on or off and a power indicator LED with a 33ω current-limiting resistor. It connects to an external host computer via a USB cable to send the collected data for tallying and recording, and to receive commands from the host computer. The USB cable also supplies power to the device.

All the sensors are powered directly from the 5V of the host computer's USB port through the USB cable. The sensors all use a single wire to facilitate communication with the microcontroller. The DHT22 module uses a

proprietary single-wire communication protocol. All the other sensors provide a continuous voltage signal proportional to the concentration of the gas they sense. These sensors are connected to the ESP32 pins with Analog-to-Digital Converters (ADCs) that convert the continuous analog voltage signal into a digital signal. The pump is also powered from the same 5V source through an IRF3205 N-Channel MOSFET. The microcontroller switches the MOSFET to operate the air pump.

4. AI Integration with the potable aflatoxin sensor platform

Approximately 5 µL of the standard solution for each sample was spotted on Whatman grade 1 paper and analyzed immediately using a developed portable AI-powered sensor platform (as shown in Figure 4). Figure 4 shows the procedure followed for data collection from the portable IoT sensor developed for processing, and the workflow of the integration of the AI machine learning algorithms (SVM, Random forest, and ANN) with feature selection and parameter tuning was implemented.

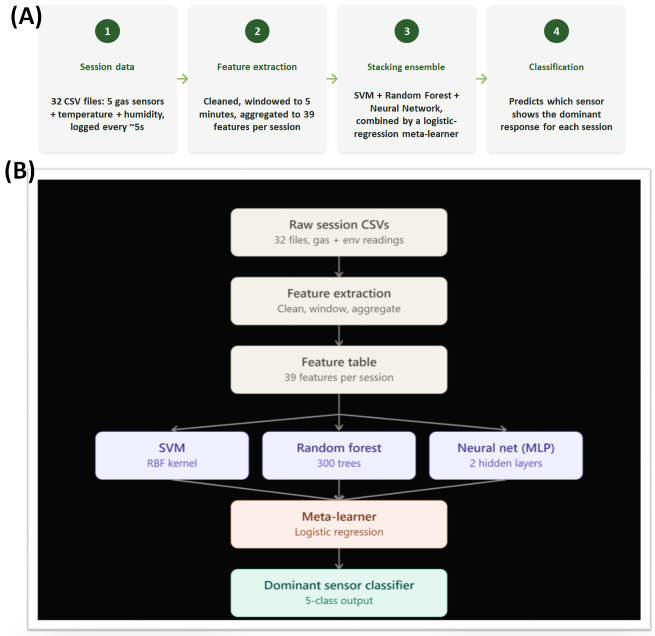


Figure 4. AI Integration with the portable aflatoxin sensor platform; (A) shows the data collection and processing pipeline of the IoT sensor platform for the Aflatoxin analysis in the open air; (B) shows the actual AI ML integration of the assembly ML algorithm developed for the Aflatoxin analysis.

III. RESULTS AND DISCUSSION

To assess the potential of the developed portable IoT sensor to analyze aflatoxin volatile compounds (VOCs) in the open air and to establish an analytical procedure, standard model VOCs were first analyzed. After establishing the analytical approach, 1-Octen-3-ol ("Mushroom Alcohol"): A very common fungal VOC associated with moldy odors and emitted by Aspergillus species [5]. It is used as a biomarker for mold contamination in grains, often with a low odor detection threshold; 3-Methyl-1-butanol and 2-Methyl-1-propanol: These are alcohols frequently associated with both aflatoxigenic and non-aflatoxigenic A. flavus strains. They are identified as key volatiles in contaminated pistachios and corn. 2-Pentylfuran: This

compound is often found in the headspace of Aspergillus flavus-contaminated grain and is highly associated with A. flavus infection in samples such as corn. Dimethyl Disulfide: Along with dimethyl sulfide, this compound is part of the VOC profile found in Aspergillus species-infected samples [19].

Direct Aflatoxin analysis using the developed smart IoT-based sensor. Figure 4 shows the recorded s spectra of 1-Octen-3-ol (MQ 3), 3-Methyl-1-butanol (MQ 5), 2-Pentylfuran (MQ 31), 2-Methyl-1-propanol (MQ 35), and Dimethyl Disulfide (MP 503) model compounds dissolved in [MeOH] v/v, at 1000 ppm as the standard solvent. Approx. 5 µL were spotted on Whatman chromatographic paper 1 and analyzed for 5 minutes using the developed portable IoT sensor in open air. In Figure 5(A), a stable signal is observed for all target VOCs of the Aflatoxins, with no background subtraction. Figure B shows the stable mean signal response for the Aflatoxin model VOCs detected by the developed sensor. It is very important to note that the developed device is capable of measuring both the temperature and the humidity of the environment, as shown in both Figure 5 (A and B). These results are in good agreement with results obtained previously [20]. The recordings with different timestamps were stored as a CSV file, ready for AI machine learning modeling and visualization.

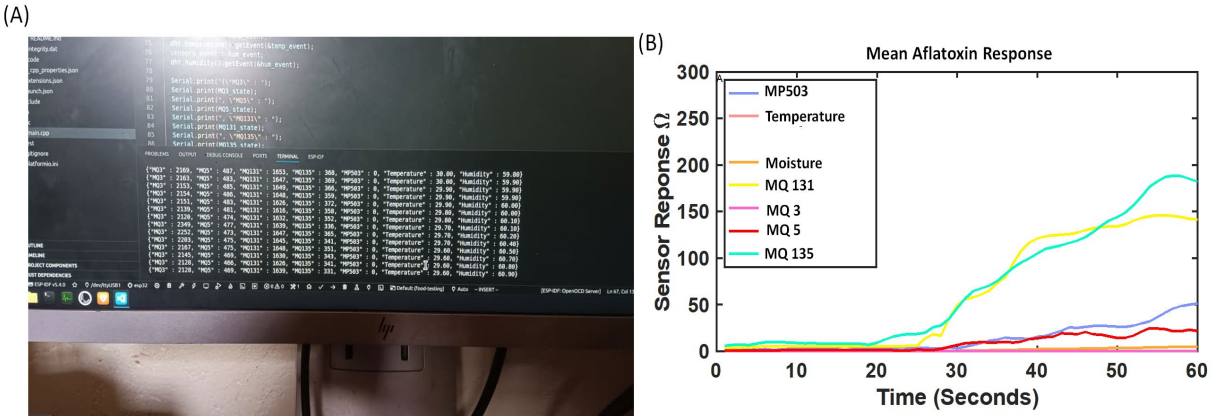


Figure 5. Analysis of Aflatoxins model VOCs using the developed IoT Sensor for 1-Octen-3-ol (MQ 3), 3-Methyl-1-butanol (MQ 5), 2-Pentylfuran (MQ 31), 2-Methyl-1-propanol (MQ 35), and Dimethyl Disulfide (MP 503) model compounds dissolved in [MeOH] v/v, at 1000 ppm. Approx. 5 µL were spotted on Whatman chromatographic paper 1 and analyzed for 5 minutes using the developed portable IoT sensor in open air. (A) and (B) show the recorded data and the mean Aflatoxin response, respectively.

AI Integration with the portable aflatoxin sensor platform. The developed IoT smart sensor was used to record data for 6 minutes and stored as CSV for the AI modelling as shown in Figure 3 Figure 6 (A to D), shows the capability of the developed sensor to detect individual a Aflatoxins VOCs; (A) shows which sensor is more dominant, (B) shows the confusion matrix obtained, while (C) and (D) shows the stacking classifier evaluated on a held out 20% validation split of the recorded data and the per class precision, recall and F1 score respectively. Figure 6(A) shows the dominant sensor; the reading that deviates most from its own baseline (z-score) per session. For the 31 labeled sessions, 85.7% accuracy and Cohen's kappa of 0.78 were obtained on real sensor data. The system performance was validated using the Stacking classifier on a held-out 20% validation split. Figure 6(B) shows the model's confusion matrix, with one error: the MQ135 session was predicted as MQ5-dominant.

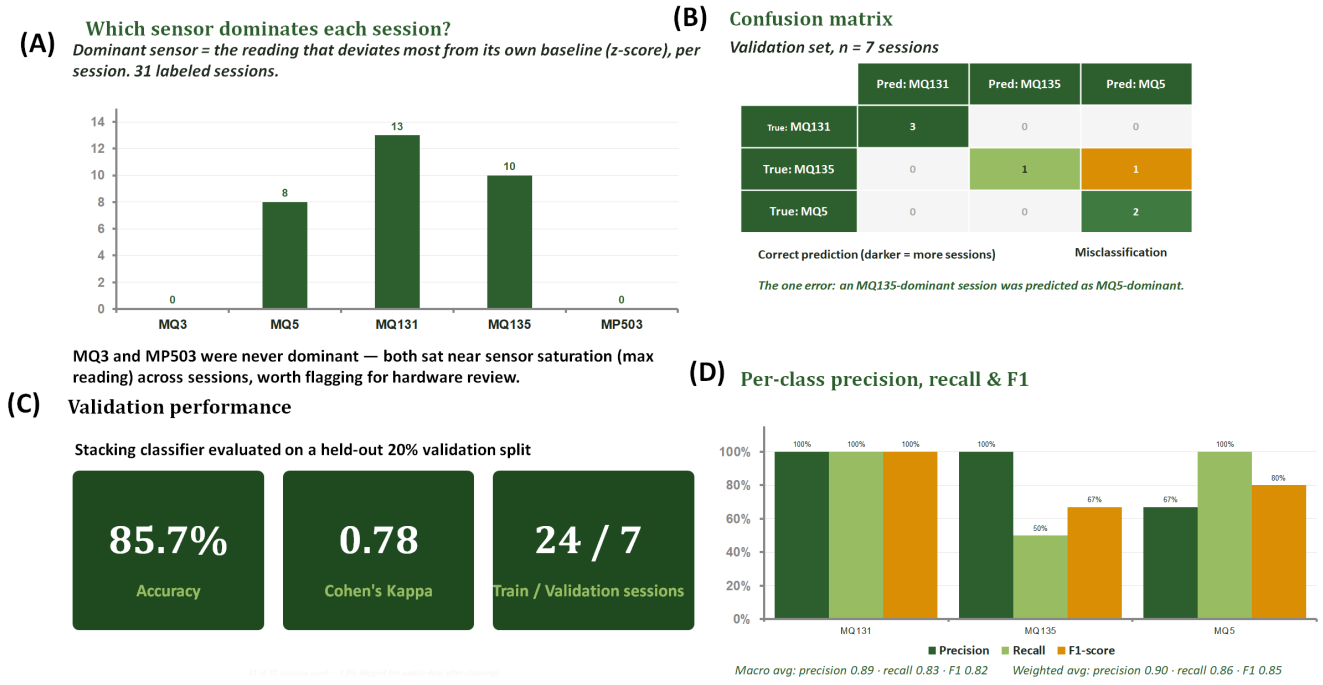


Figure 6. AI model for the detection of Aflatoxins detected using the developed IoT smart sensor; (A to D), shows the capability of the developed sensor to detect individual a Aflatoxins VOCs; (A) shows which sensor is more dominant, (B) shows the confusion matrix obtained, while (C) and (D) shows the stacking classifier evaluated on a held out 20% validation split of the recorded data and the per class precision, recall and F1 score respectively.

In this study, an AI-powered IoT smart sensor platform for rapid, portable, and low-cost detection of aflatoxin using VOCs specific to the aflatoxin species is demonstrated. The developed platform integrates an array of sensors within an IoT architecture and an ensemble machine learning framework that combines Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) classifiers. Unlike conventional laboratory-based techniques, the proposed system can allow rapid in-field detection with minimal sample preparation while maintaining high analytical performance. The main contributions of this work are summarized as follows: (i) design and development of an AI-enabled portable IoT sensor platform for rapid detection of aflatoxin in the open environment with minimal sample preparation; (ii) development of an ensemble machine learning model integrating RF, SVM, and ANN classifiers to improve detection accuracy, robustness, and generalization with high accuracy and precision; (iii) implementation of real-time edge intelligence on an IoT platform for low-latency field deployment without dependence on cloud computing; (iv) comprehensive experimental validation using VOCs aflatoxin makers; (vi) demonstrated a scalable intelligent sensing framework capable of supporting food safety surveillance, precision agriculture, and early warning systems in resource-limited environments.

The developed platform contributes toward achieving Sustainable Development Goals (SDGs) 2 (Zero Hunger), 3 (Good Health and Well-being), and 9 (Industry, Innovation and Infrastructure), while supporting Uganda's Vision 2040 and National Development Plan IV through the development of affordable intelligent food safety technologies. It's very important to note that the experiments were conducted using VOC standards deposited on filter paper rather than naturally contaminated maize, peanuts, or grain samples. Real agricultural samples

contain moisture, dust, a mix of VOCs, fungi, and environmental interference. Therefore, laboratory performance may not reflect real-world performance, which is the main focus of future work.

IV. CONCLUSION

Centralized laboratory-based detection methods for Aflatoxins, such as ELISA and HPLC, are expensive and time-consuming and are also constrained by limited laboratory capacity and sample transportation challenges in developing countries. This bottleneck prevents timely detection and intervention, leaving producers and consumers exposed to the health risks of Aflatoxins. In this study, the ability to detect and characterize Aflatoxins has been demonstrated by detecting the active, species-specific volatile compounds using a portable, smart, IoT-based sensor. The use of the sensor array to detect VoCs specific to a class of Aflatoxins is novel and shows promise; however, the stability of the sensors used must be investigated further with different VOCs and validated using a laboratory-based instrument such as a GC-MS or HPLC with tandem mass spectrometry. Further studies could address this limitation and/or investigate the possibility of Aflatoxin analysis using a high-resolution mass spectrometer. With high accuracy and precision of 85% achieved using AI machine learning, the method demonstrated in this study shows great promise for the holistic monitoring of food products, with little or no sample preparation. The results obtained indicate that the portable AI powered sensor for aflatoxins is a green sensor for field applications, without the use of solvents or gases for the analysis of analyte(s) in neat solution near real-time analysis and quantification using assemble machine learning AI model that combines random forest, SVM and ANN with 85% to 100% accuracy and sensitivity obtained for the analysis of Aflatoxins in the open environment with little to no sample preparation.

ACKNOWLEDGEMENTS

Funding for this research has been provided by the Government of Uganda, Mbarara University of Science and Technology (MUST), and the Faculty of Applied Sciences and Technology (FAST).

REFERENCES

[1] A. N. Mafe and D. Büsselberg, "Mycotoxins in food: cancer risks and strategies for control," Foods, vol. 13, no. 21, p. 3502, 2024.
[2] A. A. Goda et al., "Global health and economic impacts of mycotoxins: a comprehensive review," Environmental Sciences Europe, vol. 37, no. 1, p. 122, 2025.
[3] M. T. Deshmukh et al., "Towards intelligent food safety: Machine learning approaches for aflatoxin detection and risk prediction," Trends in Food Science & Technology, vol. 161, p. 105055, 2025.
[4] T. K. Patel, "Mycotoxins: Challenges and strategies for mitigation," in Mycological inventions for sustainable agriculture and food production: IGI Global Scientific Publishing, 2025, pp. 377-406.
[5] P. Wacoo, "Detection and reduction of aflatoxin in the food chain," 2020.
[6] T. Omara et al., "The Scourge of Aflatoxins in Kenya: A 60-Year Review (1960 to 2020)," Journal of Food Quality, vol. 2021, no. 1, p. 8899839, 2021.
[7] K. S. Magembe, "Mycotoxins impact in food, human and animal health with special reference to aflatoxins, fumonisins, ochratoxins, zearalenone, and deoxynivalenol: A 13," European Journal of Research in Medical Sciences, vol. 10, no. 1, p. 2563, 2025.
[8] D. Negash, "A review of aflatoxin: occurrence, prevention, and gaps in both food and feed safety," Journal of Applied Microbiological Research, vol. 1, no. 1, pp. 35-43, 2018.

[9] E. Yohannis et al., "Aflatoxin contamination in agri-food systems: a comprehensive review of toxicity, food security, economic impacts, and sustainable mitigation across the value chain," Food Science & Nutrition, vol. 13, no. 10, p. e71104, 2025.
[10] A. Jallow, H. Xie, X. Tang, Z. Qi, and P. Li, "Worldwide aflatoxin contamination of agricultural products and foods: From occurrence to control," Comprehensive reviews in food science and food safety, vol. 20, no. 3, pp. 2332-2381, 2021.
[11] H. Peng, Y. Chang, R. C. Baker, and G. Zhang, "Interference of mycotoxin binders with ELISA, HPLC, and LC-MS/MS analysis of aflatoxins in maize and maize gluten," Food Additives & Contaminants: Part A, vol. 37, no. 3, pp. 496-506, 2020.
[12] S. N. Karim et al., "Innovative technological approaches in the detection and mitigation of food toxicants," Food Reviews International, vol. 42, no. 4, pp. 2309-2353, 2026.
[13] R. Köppen, M. Koch, D. Siegel, S. Merkel, R. Maul, and I. Nehls, "Determination of mycotoxins in foods: current state of analytical methods and limitations," Applied microbiology and biotechnology, vol. 86, no. 6, pp. 1595-1612, 2010.
[14] S. Deng, X. Wu, Y. Shi, H. S. El-Mesery, and X. Zhang, "Uncovering Analytical Patterns for Hazardous Components in Agricultural Production Systems," Foods, vol. 14, no. 18, p. 3261, 2025.
[15] S. Adeyeye and B. Adeyeye, "Applications of artificial intelligence, internet of things and blockchain in prevention, control and detection of mycotoxins in cereals and legumes," World Mycotoxin Journal, vol. 1, no. aop, pp. 1-36, 2026.
[16] J. O. Akullo et al., "A comprehensive review of aflatoxin in groundnut and maize products in Africa: Prevalence, detection and mitigation strategies," Journal of Food Quality, vol. 2025, no. 1, p. 2810946, 2025.
[17] M. Georgiadou et al., "Volatile profiles of healthy and aflatoxin contaminated pistachios," Food Research International, vol. 74, pp. 89-96, 2015.
[18] D. Agbetiameh et al., "Potential of atoxigenic Aspergillus flavus vegetative compatibility groups associated with maize and groundnut in Ghana as biocontrol agents for aflatoxin management," Frontiers in Microbiology, vol. 10, p. 2069, 2019.
[19] J. W. Bennett and A. A. Inamdar, "Are some fungal volatile organic compounds (VOCs) mycotoxins?" Toxins, vol. 7, no. 9, pp. 3785-3804, 2015.
[20] A. Ahmad et al., "A smart sensor system for gas fingerprint classification in Cistus Ladanifer essential oil," Journal of Food Measurement and Characterization, vol. 20, no. 4, pp. 6213-6241, 2026.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Neuro-Symbolic Correction of Kazakh Word Forms through Morphological Analysis and Selective Abstention

Laura Baitenova
Department of Information Technology, Turan University
Almaty, Kazakhstan
l.baitenova@turan-edu.kz
ORCID: 0000-0002-1591-2235

Gauhar Munaitbas
Home Credit Bank JSC
Almaty, Kazakhstan
gmunaitbas@gmail.com
ORCID: 0009-0008-7142-5723

Akhmed Baikhojayev
School of Digital Technologies, Narxoz University,
Almaty, Kazakhstan;
akhmed.baikhodzhayev@narxoz.kz
ORCID: 0009-0005-5757-0519

Gulnar Mukhamejanova
School of Digital Technologies, Narxoz University
Almaty, Kazakhstan
gulnar.mukhamedzhanova@ narxoz.kz
ORCID: 0000-0003-3368-9927

Saule Tussupova
Department of Information Technology, Turan University
Almaty, Kazakhstan
s.tussupova@turan-edu.kz
ORCID: 0000-0001-8280-1837

Daryn Nurtaza
Department of Information Technology, Turan University,
Almaty, Kazakhstan
25262929@turan-edu.kz

Abstract
The correction of Kazakh word forms is complicated by agglutinative structure, morphological ambiguity, and affix allomorphy governed by phonological rules. This paper reports the design and an initial implementation of a neuro-symbolic architecture that combines formal morphological analysis, morphology-constrained candidate generation, and a three-way decision rule that may abstain rather than correct. The prototype performs script normalization, reverse-trie analysis constrained by vowel harmony and allomorph selection, edit-based candidate generation with stem recombination, and confidence-gated routing of each token to KEEP, CORRECT, or REVIEW. We evaluate it on the UD Kazakh-KTB treebank and on a controlled error corpus generated inside each cross-validation fold. The analyzer covers 99.7% of content tokens, and 62.7% admit more than one parse. Candidate generation achieves Recall@1 of 26.2%, with the gold form retrieved; it ranks first in 78.3% of cases, so generation, rather than ranking, is the limiting stage. The three-way rule raises correction precision to 49.7% against 14.2% for a binary corrector, a difference significant under paired testing, and issues no automatic corrections on unmodified normative text, against a false-correction rate of 1.9% for the binary corrector. It does not improve aggregate correction quality: recall is 2.5%, and F0.5 is indistinguishable from the binary baseline. Two methodological findings are reported. Excluding wrongly replaced tokens from the recall denominator inflates the recall of systems that intervene frequently. A frequency dictionary built without fold isolation inflated F0.5 by 7.3 points and reversed the sign of the baseline comparison.

Keywords: automatic text correction, Kazakh language, neuro-symbolic model, morphological analysis, selective abstention, false-correction rate

I. INTRODUCTION

Automatic text correction in Kazakh faces difficulties due to agglutination and vowel harmony: a single lemma can yield many grammatically correct forms, and an identical character sequence can correspond to different morphological structures. The surface appearance of a word form is therefore not a sufficient criterion for error detection; the acceptability of its morphological structure and its fit to the context must both be checked.

Early correction systems relied on dictionaries, noisy-channel models, and the Damerau–Levenshtein distance [1], [2], and handled simple typographical errors when the correct form was present in the lexicon. Corpus and frequency models extended coverage through automatically extracted word forms and typical substitutions [3]. Still, frequency does not guarantee normativity: a common error may outrank a rare term or a correctly formed derivative. Character-level comparison also fails to separate candidates that are equidistant from the input but differ in grammatical function. A hybrid architecture for Kazakh morphological processing that combines formal, statistical, and transformative components is presented in [4]. Formal morphological analyzers [5], [6] prune structurally impossible variants by checking the admissibility of transitions between morphological states under phonological constraints, but they do not always yield an unambiguous analysis. Neural models built on symbolic, morphemic, and contextual representations [7–11] capture such dependencies better, yet with limited labeled data, they may learn spurious statistical correlations rather than productive rules, and propose contextually plausible but morphologically unacceptable substitutions. The research gap lies not in the absence of individual components but in the insufficient elaboration of their coordinated interaction: morphology should constrain the candidate space, competing interpretations should survive until contextual processing, and the decision should remain transparent. The risk of falsely correcting an already valid word from a name, a term, or a rare derivative is often underestimated. However, its practical cost may exceed that of retaining a doubtful token. The contribution of this report is fourfold. First, we describe and evaluate an implemented correction pipeline in which morphological structure constrains candidate generation rather than filtering neural output after the fact. Second, we treat the false correction of a valid word form as an independent evaluation target and measure it directly; to our knowledge, no prior work on Kazakh correction reports this quantity. Third, we identify two evaluation pitfalls that materially alter conclusions for corpora of this size: an underspecified recall denominator and frequency-dictionary leakage across cross-validation folds, and quantify both. Fourth, we specify the components of the full architecture that remain unimplemented, together with the measurements that motivate them, so that the extended system can be compared against the present baseline.

II. RELATED RESEARCH

1. Correction in morphologically rich languages

Correction systems for morphologically rich languages have combined edit-based candidate generation with morphological validation and contextual ranking. Oflazer and Guzey [12] demonstrated the value of morphological processing for agglutinative languages; Sorokin [13] showed that character-level similarity must be complemented by morphological and contextual evidence; Rozovskaya and Roth [14] showed that large normative corpora partly compensate for limited error-annotated data. For Turkish, Kara et al. [15] built a rule-driven synthetic generation pipeline that covers more than twenty expert-curated writing rules and released parallel data along with detection and correction baselines, confirming that controlled corruption is a viable substitute for scarce annotated error data in agglutinative languages. Across these studies, candidate generation and linguistic verification generally remain separate stages.

2. Morphological synthesis with limited data

Reflection studies show that formal constraints, neural generation, transfer learning, and synthetic data support word-form generation in low-resource settings [16–18]. These methods assume that the lemma and grammatical features are known. In correction, they must first be recovered from the erroneous token and its context, after which the generated form must be verified by reverse analysis.

3. Contextual morphological disambiguation for the Kazakh

Disambiguation models for Kazakh and related languages encode alternative analyses together with sentence context and select the most compatible interpretation [19], [20]. Hybrid architectures have demonstrated the complementarity of formal analyzers and contextual language models in Kazakh morphological processing [4]. These approaches improve the selection of an analysis but do not themselves construct and verify a corrected surface form.

4. Selective prediction and correction safety

Systems that may abstain rather than decide are studied under selective prediction, where performance is characterized by a risk–coverage curve rather than a single accuracy figure. Traub et al. [21] analyze the evaluation of such systems and show that conclusions drawn at a single operating point can be misleading, thereby motivating the reporting of the full curve. We adopt this framing: the REVIEW decision is an abstention, and the false-correction rate is the risk quantity it is designed to control.

5. Selective prediction and correction safety

The proposed method connects correction-candidate generation, morphological verification, and thresholded decision-making. The formal analyzer constrains admissible morphological paths, the decision rule separates confident corrections from uncertain cases, and uncertain cases are retained for review rather than corrected automatically.

III. MATERIALS AND METHODS

1. Notation and hyperparameter values

Table 1. Notation and calibrated values.

Symbol	Value	Meaning
δ	2	maximum Damerau–Levenshtein distance
κ_min	0.40	morphological compatibility floor
τ_e	0.05	error-probability gate
τ_c	0.10	top-1 probability floor for CORRECT
τ_m	0.02	top-1 vs top-2 margin floor (not binding)
λ_ed	1.0	edit-distance energy weight
λ_morph	0.5	residual morphological uncertainty weight

λ_freq	0.3	corpus-frequency weight
λ_ctx	0.0	contextual weight (no trained ranker)
ρ_stem	0.15	energy penalty for stem-recombined candidates
K	1, 5, 10	Recall@K cutoffs

Thresholds τ_e, τ_c, and τ_m were calibrated on fold 0 and fixed before folds 1–4 were evaluated. Fold 0 does not appear in any reported result.

2. Processing pipeline

The system takes a token sequence and returns a sequence of the same length, with each token either retained or replaced; ambiguous cases are marked for manual review rather than replaced automatically. The processing order is: orthographic and script normalization, morphological analysis, error estimation, candidate generation, candidate ranking, and one of three final decisions: KEEP, CORRECT, or REVIEW.

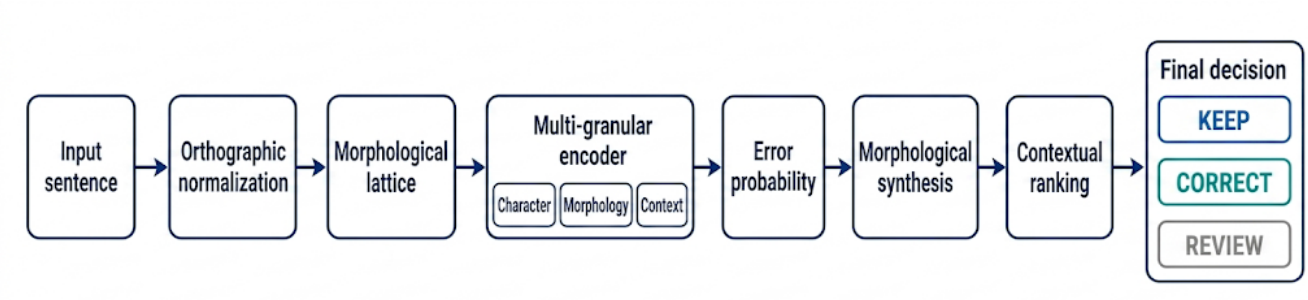


Figure 1. Overall architecture of the neuro-symbolic correction pipeline.

3. Language resources

Table 2. Data resources and their function.

Resource	Size	Role
UD Kazakh-KTB [22]	1,078 sentences; 10,536 tokens	analysis evaluation, corruption source, correction evaluation
Apertium-kaz lemma inventory	≈22,000 lemmas	candidate index (SymSpell, radius δ = 2)
Controlled error set	862 pairs/fold-set (folds 1–4), regenerated × 3 seeds (≈65% corrupted, ≈35% identity); E = 1,683 injected-error tokens per evaluation point	error detection, candidate generation, correction

All texts were converted to a uniform Unicode representation; spaces, quotation marks, hyphens, and apostrophes were normalized. Kazakh–Russian grapheme substitutions were not silently corrected during preprocessing, since mixed-script forms constitute a separate error category.

The lexical representation of a word form w is

$$\mathrm{Lex}(w) = \{(\ell, m, p, f, \nu)\} \quad (1)$$

where ℓ is the lemma, m the morpheme sequence, p the universal part of speech, f the feature bundle, and ν the frequency estimated from the training partition. Frequency is used only as a secondary ranking signal and never as evidence of grammatical validity.

The corpus frequency dictionary is rebuilt independently for each fold from training sentences only. An earlier version of the pipeline was built once over the whole treebank; the effect of that overlap is quantified in Section IV-F.

4. Controlled error generation

Synthetic distortions are introduced only after document-level splitting. For a normative sentence Y, the corruption operator generates an input sentence

$$X = q_{\phi}(Y, t, k) \quad (2)$$

where t is the error type, k the token position, and φ the operation-specific parameters. Each corrupted–clean pair is stored with the token position, original form, corrupted form, error category, and the morphological analysis of the normative token. Error-free identity examples are retained to estimate the false-correction rate and to prevent the model from learning that every sentence contains a replacement.

Table 3. Controlled corruption operators.

Error category	**Corruption operation**	**Status**
Character deletion	remove one grapheme	implemented
Character insertion	insert a keyboard-neighbor or sampled grapheme	implemented
Character substitution	Replace a grapheme with a visually or keyboard-related one	implemented
Adjacent transposition	exchange two neighboring graphemes	implemented
Mixed-script substitution	Replace a Kazakh grapheme with a visually similar Russian Cyrillic one	implemented
Vowel-harmony violation	replace an affix with an incompatible allomorph	implemented
Affix-vowel deletion	remove a vowel from an affix	implemented

Grammatical-affix substitution	replace case, possessive, person, or agreement marking	implemented
Identity example	preserve the normative sentence unmodified	implemented

Each operation is applied only when its linguistic preconditions are satisfied. Examples producing another contextually valid word form are excluded or assigned to a separate ambiguity category.

5. Morphological analysis

For each token the analyzer produces a set of admissible interpretations

$$\mathcal{A}_i = \{a_{ij}\}, \quad j = 1 \dots M_i, \quad a_{ij} = \left(\ell_{ij}, m_{ij}, p_{ij}, f_{ij}\right) \tag{3}$$

retaining paths that satisfy stem compatibility, morphotactic order, vowel harmony, and allomorph selection. The implementation is a reverse-trie suffix matcher over 67 affix classes with 64 harmony and allomorph-selection rules, following the finite-state principle used in previous Kazakh analyzers [5], [6].

Specified but not implemented. The full architecture preserves M_i > 1 competing paths in a lattice until contextual disambiguation. The present implementation returns a single analysis per token. The measured proportion of tokens admitting more than one path (Section IV-C) quantifies what this simplification discards.

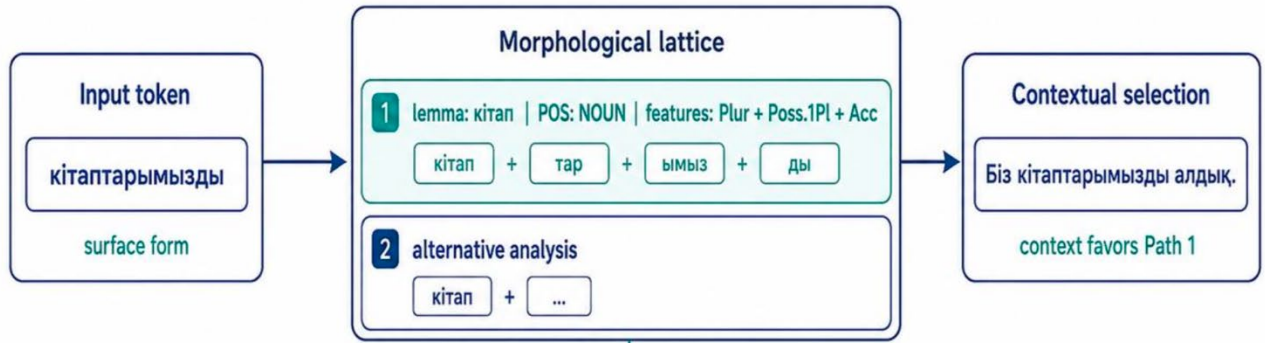


Figure 2. Morphological analysis of a Kazakh word form.

6. Candidate generation

The specified candidate set requires a morphological synthesizer:

$$\mathcal{C}_i^{\text{spec}} = \{\, c : c = \text{Synth}(\ell', f'),\ d(x_i, c) \le \delta,\ \text{Analyze}(c) \supseteq (\ell', f') \,\} \tag{4}$$

This condition specifies the intended verification step. The present implementation approximates it on the analysis side:

$$\mathcal{C}_i = \{\, c : d(x_i, c) \le \delta,\ \text{Analyze}(c) = (\ell', m'),\ \text{Reattach}(\ell', m') = c,\ \kappa(c) \ge \kappa_{\min} \,\} \qquad (5)$$

A candidate is retained if its stem is recovered by the analyzer, if re-attaching the recovered suffix sequence reproduces the candidate surface form, and if the re-parsed candidate attains compatibility κ ≥ κ_min. Candidates are additionally produced by recombining a recovered stem with admissible affix chains, at an energy penalty ρ_stem.

Specified but not implemented. A morphological synthesizer generating a surface form from a lemma and feature bundle, as in [6], is not implemented. The per-category results in Section IV-D show the consequence directly.

7. Ranking and the three-way decision rule

Each candidate receives an energy score $Büyük E(c \mid X, i) = -\lambda_{\text{ctx}} \log P_{\text{ctx}}(c \mid X, i) + \lambda_{\text{ed}}\, \tilde{d}(x_i, c) + \lambda_{\text{morph}}\, \Omega(c) - \lambda_{\text{freq}} \log(1 + \nu_c)$ (6)

where $\tilde{d}$ is the edit distance normalized by the length of the input token, Ω(c) is the residual morphological uncertainty of the candidate, and ν_c its training-partition frequency. Morphologically invalid candidates are removed before this score is computed. Normalised candidate probabilities are $P(c \mid X, i) = \frac{\exp[-E(c|X,i)]}{\sum_{c' \in \mathcal{C}_i \cup \{x_i\}} \exp[-E(c'|X,i)]}$ (7)

Sağriginal token is included in the same decision set as the candidates, making preservation a first-class decision rather than a fallback. c, son taban, alt 2, the two highest-ranked alternatives, the confidence margin is büyük harf delta alt i. EşittirWith c_1 and , c_2 the two highest-ranked alternatives, the confidence margin is $\Delta_i = P(c_1 \mid X, i) - P(c_2 \mid X, i)$, and the decision rule is

$$D_i = \begin{cases} \text{KEEP}, & p_i^{\text{err}} < \tau_e \ \text{ or } \ c_1 = x_i, \\ \text{CORRECT}, & p_i^{\text{err}} \ge \tau_e,\ P(c_1 \mid X, i) \ge \tau_c,\ \Delta_i \ge \tau_m, \\ \text{REVIEW}, & \text{otherwise.} \end{cases} \qquad (8)$$

All three conditions are active at the calibrated operating point. Proper names, abbreviations, and unknown stems are assigned a stricter automatic-correction threshold. REVIEW preserves the original token and returns an ordered set of candidates with morphological analyses.

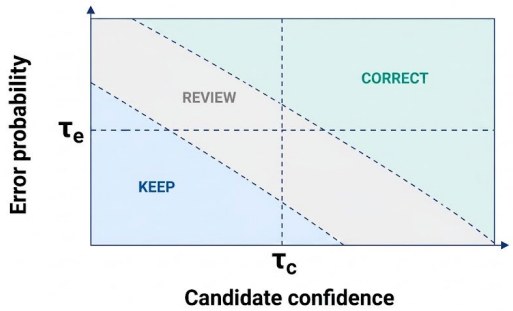


Figure 3. Three-way decision regions in the (error probability, candidate confidence) plane.

8. Specified neural components

Three components of the full architecture are specified but not implemented, and no equations for them are provided here, as none were trained.

A multi-granular representation would combine character-level, morphology-level, and sentence-context encodings through a learned gate, following the character-aware and multi-granularity modeling principles established for agglutinative languages [7], [8], [10], [11]. A joint training objective would optimize error detection, analysis selection, lemmatization, tagging, feature prediction, candidate ranking, and confidence calibration. The calibration term in that objective is intended to align model confidence with observed correctness and to support a defensible REVIEW threshold; Section IV-H shows the behavior that results when it is absent.

In the reported system, λ_ctx = 0: no contextual ranker is trained, and candidate scoring reduces to edit distance, morphological compatibility, and corpus frequency. KazBERT contributes part-of-speech confidence to the error-probability estimate but not to candidate scoring.

9. Baselines and ablations

Three systems are evaluated under identical partitions, preprocessing, and thresholds: frequency-ranked top-1 correction, a binary corrector without REVIEW routing, and the proposed three-way system. A no-op configuration, in which the input is returned unchanged, is included as a floor for the character- and word-error rates. Character-only and context-only neural baselines are specified in the comparison design but were not evaluated in this report, since they require the multi-granular encoder described in Section III-H. Ablations remove stem recombination, morphological filtering, and REVIEW routing. The ablation of lattice encoding is not reported as a measured row because the component is not implemented.

10. Evaluation metrics and protocol

Candidate-generation recall at cutoff K is

$$\text{Recall@}K = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\, y_i \in \text{Top}K(\mathcal{C}_i) \,\right] \tag{9}$$

Final correction is evaluated with precision, recall, and

$$F_{0.5} = \frac{(1+0.5^2)\,P\,R}{0.5^2 P + R} \tag{10}$$

which weights precision more heavily than recall, because an unjustified change to a correct word form is more costly than retaining an uncertain token. Recall is computed against the fixed number of injected errors in the fold, so that a token replaced with an incorrect form counts as both a missed correction and a false positive; denominators are therefore identical across systems. The false-correction rate is

$$\text{FCR} = \frac{N(\text{changed correct tokens})}{N(\text{correct tokens})} \tag{11}$$

For the selective-correction regime, risk and coverage at threshold τ are

$$R(\tau) = \frac{N(\text{incorrect automatic decisions at } \tau)}{N(\text{automatic decisions at } \tau)}, \quad C(\tau) = \frac{N(\text{automatic decisions at } \tau)}{N(\text{all evaluated cases})} \tag{12}$$

Quality is assessed in four areas: candidate generation (Recall@1/5/10, MRR), final correction (precision, recall, $F_{0.5}$, CER, WER), safety (false-correction rate), and selective behavior (the risk–coverage curve for the KEEP/CORRECT/REVIEW regime).

Protocol. Evaluation uses 5-fold stratified cross-validation on UD Kazakh-KTB. Fold 0 is held out for threshold calibration and excluded from all reported results; folds 1–4 are used for testing. Three corruption seeds (13, 42, 123) are applied, yielding 12-fold × seed evaluation points. Synthetic errors are generated independently inside each fold after splitting, preventing a normative sentence from appearing in training while its corrupted derivative appears in testing. Token-level traces — original form, generated candidates, morphological analyses, scores, final decision, and grounds for rejecting each option — are retained, permitting failures to be attributed to candidate generation, morphological analysis, ranking, or decision calibration.

Determinism. All reported numbers require full seeding of the Python, NumPy, and PyTorch random number generators, along with a fixed PYTHONHASHSEED set before interpreter startup. Without the latter, the dictionary iteration order during model and tokenizer loading varies across processes, which changes the classification of borderline tokens and shifts the reported precision by more than 10 percentage points. This sensitivity is a consequence of the low intervention rate reported in Section IV-E, with fewer than 100 automatic corrections per evaluation point; each borderline decision reduces precision by roughly 1 percentage point. Under full seeding, the pipeline produces bit-identical results across processes and across CPU and GPU backends.

Implementation details. The contextual encoder is KazBERT (Eraly-ml/KazBERT), a monolingual Kazakh BERT. Part-of-speech and lemma checkpoints were restored from a companion morphological-tagging project and are used for inference only; no network is trained in this report. Evaluation was conducted on an Apple MacBook Pro 2023 M3 Pro (MRX33) using the seeding procedure described above.

IV. RESULTS

1. Implemented and specified components

Table 4. Status of architecture components.

Component	Status
Script normalization; reverse-trie analyzer (67 affix classes); harmony and allomorph rules (64); error taxonomy	implemented
Edit-based candidate generation (SymSpell, $\delta \leq 2$)	implemented
Stem recombination with κ-gated re-parsing, eq. (5)	implemented (analysis-side approximation of eq. (4))
KEEP / CORRECT / REVIEW decision rule, eq. (8)	implemented
Morphological lattice preserving competing analyses	specified

Morphological synthesizer and analysis–synthesis–reanalysis cycle, eq. (4)	specified
Multi-granular encoder with learned gating	specified
Joint training objective with calibration term	specified

One element of the specified scoring function is inactive: the contextual energy weight λ_ctx is zero because no contextual ranker is trained; thus, candidate ranking uses only edit distance, morphological compatibility, and corpus frequency. The margin threshold τ_m is calibrated to 0.02 and is binding (a candidate must lead by ≥0.02 probability to be auto-applied). This is reported as a measured configuration setting and is revisited in the Limitations.

2. Experimental setup

UD Kazakh-KTB, 1,078 sentences and 10,536 tokens; 5-fold stratified cross-validation with fold 0 held out for calibration; three corruption seeds; corruption applied inside each fold after splitting; frequency dictionary rebuilt per fold from training sentences only.

3. Analyzer coverage

The analyzer produces an admissible analysis for 99.7% of the 8,437 content tokens (8,415 tokens, excluding punctuation and symbols) and recovers the correct lemma for 96.4% of tokens (8,131 tokens). Coverage is uniform across parts of speech, ranging from 93.9% (PART, n = 33) to 100.0%; the next lowest values are PROPN (98.6%, n = 562) and AUX (98.6%, n = 361).

62.7% of content tokens (5,290) admit two or more admissible suffix paths. This is the empirical justification for the lattice representation specified in Section III-E. For the majority of content tokens, more than one morphological parse is admissible, and collapsing to a single analysis before contextual disambiguation discards viable interpretations.

4. Candidate generation

Table 5. Candidate generation on the controlled error set (5 folds × 3 seeds, leakage-free protocol).

Configuration	**R@1**	**R@5**	**R@10**	**MRR**
DL, $\delta \leq 2$	19.6 ± 4.0	26.2 ± 4.6	28.0 ± 4.9	22.4 ± 4.1
DL + freq	22.2 ± 3.8	27.4 ± 4.6	28.6 ± 4.2	24.5 ± 4.1
DL + morph filter	19.8 ± 4.1	26.4 ± 4.6	28.1 ± 4.8	22.6 ± 4.1
Proposed	26.2 ± 4.0	32.1 ± 4.5	33.4 ± 4.1	28.8 ± 4.2

The Recall@1/Recall@10 ratio is 78.3% for the proposed generator: once the gold form is retrieved, it almost always ranks first. Retrieval rather than ranking is therefore the limiting stage — 20.9% of corrupted tokens yield no candidate at all, and Recall@10 saturates at 33.4%

Table 6. Recall@K by corruption category (proposed generator, fold-averaged, seed 0).

Category	N	R@1	R@5	R@10
Character insertion	134	50.7	51.5	51.5
Mixed-script substitution	79	35.4	38.0	38.0
Character substitution	112	33.9	38.4	38.4
Character deletion	141	27.7	40.4	41.1
Adjacent transposition	127	18.1	31.5	33.9
Vowel-harmony violation	44	11.4	13.6	13.6
Affix-vowel deletion	60	1.7	3.3	5.0
Grammatical-affix substitution	23	0.0	0.0	0.0

The breakdown localizes the failure. Character-level corruptions are recovered at 18%–51% Recall@1, whereas affix-vowel deletion (1.7%) and grammatical-affix substitution (0.0%) are essentially unreachable. These are precisely the categories that require generation from a lemma and feature bundle rather than edit-based retrieval, and they constitute the empirical case for the morphological synthesizer specified in Section III-F.

5. Correction quality

Table 7. Final correction on the controlled error set (folds 1–4 × 3 seeds; thresholds calibrated on fold 0: τ_e = 0.05, τ_c = 0.10, τ_m = 0.2). Number of injected errors E = 1,683 for all systems.

System	P	R	$F_{0.5}$	FCR↓	CER↓	WER↓
No-op	—	—	—	—	2.7 ± 0.3	9.7 ± 0.8
DL + freq top-1	10.7 ± 3.6	16.7 ± 4.2	11.5 ± 3.6	8.6 ± 3.7	5.5 ± 1.1	16.1 ± 3.2
Binary	14.2 ± 4.5	7.7 ± 2.0	11.9 ± 3.2	2.0 ± 1.5	3.7 ± 0.6	11.1 ± 1.5
Proposed	49.7 ± 18.9	2.5 ± 1.1	10.1 ± 4.3	0.2 ± 0.1	2.8 ± 0.4	9.7 ± 0.9

Confusion counts pooled over the twelve evaluation points: top-1 baseline TP = 283, FP = 2,500, FN = 1,400; binary corrector TP = 129, FP = 830, FN = 1,554; proposed system TP = 41, FP = 44, FN = 1,642.

The three systems occupy different points on the precision–recall trade-off. The top-1 baseline intervenes on 2,783 tokens and recovers 283 errors; the binary corrector intervenes on 959 and recovers 129; the three-way

rule intervenes on 85 and recovers 41. Precision therefore rises from 10.7% to 14.2% to 49.7%, while recall falls from 16.7% to 7.7% to 2.5%. Because λ_ctx = 0, candidate probability mass is spread thinly and the CORRECT branch fires rarely; the low recall is a consequence of the untrained ranker rather than of the decision rule itself.

Significance. Paired tests over the twelve-fold × seed points compare the proposed system with the binary corrector. On precision, the difference is significant: Δ = +0.355, 95% bootstrap interval [+0.248, +0.460], bootstrap *p* < 0.0001, Wilcoxon *p* = 0.0022. On $F_{0.5}$, the proposed system is not ahead: Δ = −0.018, 95% bootstrap interval [−0.046, +0.012], bootstrap *p* = 0.235, Wilcoxon *p* = 0.209. We therefore claim an improvement in correction precision and in correction safety, and explicitly do not claim an improvement in aggregate correction quality. The twelve points comprise four folds × three seeds and are not fully independent, so the *p*-values should be read as indicative. The large standard deviation in precision (±18.9) reflects the small number of automatic corrections per evaluation point — between three and eleven — although no evaluation point produced zero corrections, so the paired test is well-defined throughout.

Character and word error rates. Both quantities for the proposed system are at the no-op floor. This follows directly from the intervention rate: 41 errors repaired against 44 introduced, a net change of three errors in 1,683. We report CER and WER for completeness and do not present them as an advantage of the three-way rule. The binary corrector, by contrast, raises both rates well above the floor (3.7% and 11.1% against 2.7% and 9.7%), confirming that unconstrained intervention degrades text quality on this corpus.

6. Leakage-controlled comparison

An earlier version of the evaluation built the corpus frequency dictionary once over the entire treebank, so gold forms from the test folds were visible to the frequency-ranking modes.

Table 8. Effect of frequency-dictionary leakage (identical recall denominator; folds 1–4 × 3 seeds).

System	$F_{0.5}$ leaky	$F_{0.5}$ leakage-free	Inflation
DL + frequency, top-1	12.4 ± 4.1	11.5 ± 3.6	+1.0 pp
Binary corrector	11.9 ± 3.2	11.9 ± 3.2	0.0 pp
Three-way (proposed)	17.4 ± 6.5	10.1 ± 4.3	+7.3 pp

Protocol	Δ $F_{0.5}$	95% bootstrap CI	bootstrap *p*	Wilcoxon *p*
Leakage-free (reported)	-0.018	[−0.046, +0.012]	0.235	0.209
Leaky (original)	+0.055	[+0.027, +0.082]	< 0.0001	0.0012

Candidate Recall@1 for the proposed generator falls from 29.0% to 26.2% under fold isolation. Modes that do not consult the frequency dictionary are unaffected. The effect is not a matter of degree. Under the leaky protocol, the proposed system appears to outperform the binary baseline by a statistically significant margin; under fold isolation, the same comparison shows a small, non-significant deficit. The leakage therefore reversed the

conclusion's sign, not merely its magnitude. On a treebank of this size, a frequency resource built without fold isolation can yield spurious results, and corpus- and frequency-based correction methods for low-resource languages are commonly evaluated without this control. We report both protocols so that the comparison can be reproduced under either.

7. False correction on normative text

Table 9. False-correction rate on unmodified normative text (folds 1–4, 862 sentences, 8,270 tokens).

Configuration	FCR	KEEP	CORRECT	REVIEW
Binary corrector (− REVIEW)	1.9% (154/8,270)	8,116	154	0
Three-way (proposed)	**0.0%** (0/8,270)	7,996	0	274

On unmodified normative text, the three-way rule issued no automatic corrections at all: 7,996 tokens were kept, and 274 (3.3%) were routed to review, giving a false-correction rate of 0.0% across all four evaluation folds, compared with 1.9% for the binary corrector. This is not the trivial property of a system that never acts. On the controlled error set, the same rule, under the same thresholds, issued 85 corrections with 49.7% precision (Table 7). The rule, therefore, separates the two conditions: it intervenes when corruption is present and abstains when the input is already normative. What the clean-text measurement establishes is the absence of a failure mode, not the presence of a capability.

Per-fold false-correction rates for the binary corrector range from 0.90% to 4.38%, so the pooled figure conceals substantial variation.

Table 10. Stratified false-correction rate (binary corrector, pooled).

Partition	**N**	**FCR**
Not in the lexicon	5,870	2.5% (148)
Common	1,773	0.0% (0)
Proper name	440	1.4% (6)
Borrowing	187	0.0% (0)
Analyzer: known	8,249	1.9% (154)
Analyzer: unknown	21	0.0% (0)

The stratification must be read carefully. Tokens absent from the lemma lists account for 5,870 of 8,270 tokens and for 148 of the 154 false corrections; since this partition covers 71% of the corpus, the concentration is partly a property of the partition rather than of the errors. The analyzer-based partition is sharper: only 21 tokens are unparseable, and none received a false correction. The large lexicon-miss class, therefore, reflects a narrow frequency lexicon rather than words that the system cannot analyze.

To the authors' knowledge, no prior work on Kazakh correction reports a false-correction rate; the cited systems report only the accuracy of repairing injected errors.

8. Risk–coverage behavior

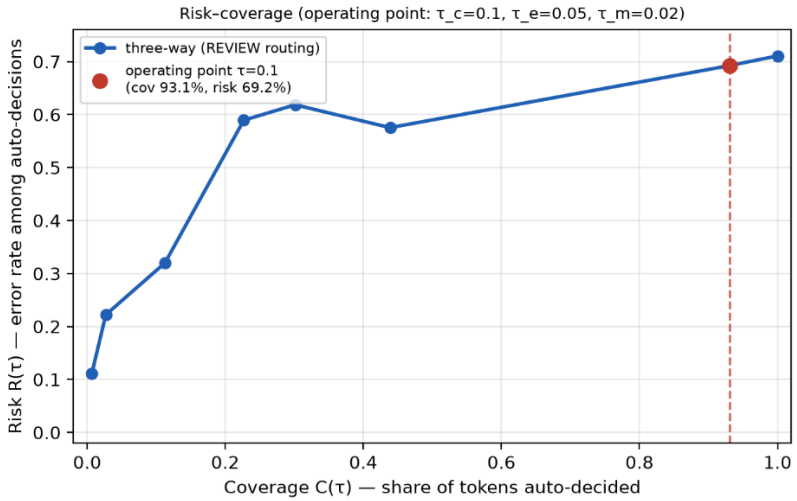


Figure 4. Risk R(τ) against coverage C(τ) for the three-way rule, with the calibrated operating point τ_c = 0.10 marked.

At the selected threshold τ_c = 0.10, the top-1 condition admits 93.1% of the 1,331 candidate-bearing tokens at 69.2% risk. Risk falls to 32.0% at 11.3% coverage, to 22.2% at 2.7% coverage, and to 11.1% at 0.7% coverage. The curve, therefore, exposes the trade-off and makes it selectable, but it does not offer a favorable operating region at any threshold that retains meaningful coverage. The curve exposes the trade-off and makes it selectable, but it does not yet offer a favorable operating region.

The curve is not monotone: risk rises from 57.5% at τ = 0.15 to 61.8% at τ = 0.20. Since λ_ctx = 0, the score entering the threshold is a normalized energy over edit distance, morphological compatibility, and frequency, and it is not calibrated against observed correctness. The calibration term specified in Section III-H is the component intended to remove this behavior, and its absence is the most direct explanation for the non-monotone region. As Traub et al. [21] note, conclusions drawn from a single working point of such a system can be misleading; we therefore report the full curve rather than the operating point alone.

9. Component ablation

Table 11. Ablation on candidate Recall@K (5 folds × 3 seeds). Each row removes one component from the proposed generator.

Configuration	R@1	R@5	R@10	MRR	Δ R@1
− stem recombination (analysis filter only)	19.8 ± 4.1	26.4 ± 4.6	28.1 ± 4.8	22.6 ± 4.1	−6.4 pp

− morphology (frequency ranking only)	22.2 ± 3.8	27.4 ± 4.6	28.6 ± 4.2	24.5 ± 4.1	−4.0 pp
Full proposed	26.2 ± 4.0	32.1 ± 4.5	33.4 ± 4.1	28.8 ± 4.2	—

The morphological re-analysis filter, applied on its own, is not effective (19.8% vs. 19.6% for plain edit-distance retrieval). The measured gain arises from stem recombination combined with frequency ranking, not from filtering out morphologically implausible candidates. The REVIEW-routing ablation is reported in Table 9 on clean text and in Table 7 on the controlled error set. The lattice-encoding ablation targets a component that is not implemented; rather than report a fabricated row, we report the quantity that motivates it — the 62.7% multi-path rate of Section IV-C.

V. DISCUSSION

The architecture treats correction as the reconstruction of a morphologically admissible word form rather than the selection of the nearest dictionary item. This distinction matters for Kazakh, where several valid forms may be equidistant from the input while expressing different grammatical functions. Morphological admissibility is enforced before ranking, so structurally impossible candidates are excluded rather than penalized probabilistically. The measurements support a narrow but specific claim. Abstention raises the precision of automatic correction 3.5-fold and removes false corrections on normative text entirely. It does not yet produce a better corrector: the system repairs 2.5% of injected errors, its aggregate $F_{0.5}$ is slightly below that of a binary corrector, and its net effect on corpus-level error rates is three additional errors in 1,683. The two components most responsible for this — a trained contextual ranker and a calibration term — are precisely the ones not yet implemented, and the risk–coverage curve directly reflects their absence. Two methodological findings are worth separating from the system itself. First, excluding incorrectly replaced tokens from the recall denominator inflates the recall of systems that intervene frequently, and the denominator must be fixed to the number of injected errors for cross-system comparison. Second, a frequency dictionary built without cross-validation fold isolation inflated the proposed system's $F_{0.5}$ by 7.3 percentage points and reversed the sign of its comparison against the binary baseline. Third, without a fixed process-level hash seed, the order of dictionary iteration during model loading altered the classification of borderline tokens and shifted measured precision by more than 10 percentage points between runs of identical code. Systems with very low intervention rates are structurally exposed to this. When a decision is taken on eighty-five tokens, each borderline case is worth more than a percentage point of precision.

VI. LIMITATIONS

Five limitations must be stated explicitly.

First, the morphological component returns a single analysis per token rather than the lattice specified in Section III-E. The measured 62.7% multi-path rate indicates the scale of what is discarded. Second, morphological synthesis and the corresponding analysis–synthesis–reanalysis verification are not implemented. The candidate generator approximates the cycle on the analysis side, and the per-category results show the consequence: affix-vowel deletion and grammatical-affix substitution are recovered at 1.7% and 0.0% Recall@1, respectively. Third, no contextual ranker is trained, so $\lambda_ctx = 0$. This is the direct cause of the low CORRECT recall and of the uncalibrated, non-monotone risk–coverage behavior. A trained contextual ranker with a calibration term is the highest-leverage next step. Fourth, the proposed system does not improve the aggregate quality of corrections. Its $F_{0.5}$ is below that of the binary corrector, though not significantly so, and its character and word

error rates are at the no-op floor. The results support claims only regarding precision and safety. Fifth, precision is estimated at between three and eleven automatic corrections per evaluation point, yielding a standard deviation of ±18.9 percentage points. The comparison against the binary corrector is significant because the gap is large, but the absolute value of 49.7% should be treated as an estimate with wide uncertainty. Sixth, evaluation uses five-fold cross-validation with three corruption seeds on a treebank of 1,078 sentences, with synthetic distortions rather than natural user errors. Synthetic errors do not reproduce the distribution of errors in natural writing, and a manually verified natural-error set is under construction.

VII. CONCLUSION

This paper reports the design and an initial implementation of a neuro-symbolic architecture for the correction of Kazakh word forms, combining formal morphological analysis, morphology-constrained candidate generation with stem recombination, and a three-way KEEP/CORRECT/REVIEW decision rule. On the UD Kazakh-KTB treebank, the analyzer covers 99.7% of content tokens, and 62.7% of these tokens admit more than one admissible parse. Candidate generation achieves Recall@1 of 26.2%, and when the gold form is retrieved, it ranks first in 78.3% of cases, identifying retrieval rather than ranking as the limiting stage. The three-way rule attains 49.7% correction precision against 14.2% for a binary corrector — a difference significant under paired testing — and issues no false corrections on unmodified normative text, at the cost of a 2.5% recall. It does not improve aggregate correction quality: recall is 2.5%, $F_{0.5}$ is slightly below the binary baseline, and character and word error rates are at the no-op floor. Two methodological findings accompany these results. Treating a wrongly replaced token as absent from the recall denominator inflates the recall of systems that intervene frequently. A frequency dictionary built without fold isolation inflated the proposed system's $F_{0.5}$ by 7.3 percentage points and reversed the sign of its comparison against the baseline. Process-level hash randomization shifted measured precision by more than 10 percentage points between runs of identical code. Both are reported so that the comparison can be reproduced under either protocol. Four components of the full architecture remain unimplemented: a morphological lattice, a synthesizer supporting analysis–synthesis–reanalysis verification, a multi-granular encoder with learned gating, and a joint training objective with a calibration term.

DATA AND CODE AVAILABILITY

All results in Section IV are reproducible from the accompanying repository *1*. Threshold calibration is performed on fold 0 and locked before folds 1–4 are evaluated; corruption is applied inside each fold after splitting; the frequency dictionary is rebuilt per fold from training sentences only. The KazBERT part-of-speech and lemma checkpoints are used for inference only. Both the leakage-free and leaky evaluation protocols are retained in the repository to enable Table 8 to be reproduced.

ACKNOWLEDGEMENT

The authors would like to acknowledge the use of large language models to assist in improving the manuscript's clarity, structure, and overall language quality. Additionally, the Grammarly tool was utilized for grammar checking and linguistic refinement. These tools were employed solely to enhance readability and presentation, while all scientific content, methodology, and results remain the authors' original work.

FUNDING STATEMENT

1 https://github.com/akhmedbike/IHCONCS_2026_paper

The authors declare that financial support was received for the research and/or publication of this article. This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. IRN AP23487753) under the project titled “Innovative technologies for automated correction of Kazakh language texts: machine learning and morphological analysis”.

REFERENCES

[1] A. Makazhanov, O. Makhambetov, I. Sabyrgaliyev, and Z. Yessenbayev, "Spelling correction for Kazakh," in Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Part II, LNCS 8404, Kathmandu, Nepal. Berlin, Germany: Springer, 2014, pp. 533–541, doi: 10.1007/978-3-642-54903-8_44.

[2] N. Mukazhanov, Z. Alibiyeva, A. Yerimbetova, A. Kassymova, and N. Alibiyeva, "Development of an augmented Damerau–Levenshtein method for correcting spelling errors in Kazakh texts," Eastern-European Journal of Enterprise Technologies, vol. 5, no. 2 (125), pp. 23–33, 2023, doi: 10.15587/1729-4061.2023.289187.

[3] A. Toleu, G. Tolegen, R. Mussabayev, A. Krassovitskiy, and I. Ualiyeva, "Data-driven approach for spellchecking and autocorrection," Symmetry, vol. 14, no. 11, Art. no. 2261, 2022, doi: 10.3390/sym14112261.

[4] L. Baitenova, S. Tussupova, S. Mambetov, G. Munaitbas, and G. Mukhamejanova, "Hybrid artificial intelligence architectures for automatic text correction in the Kazakh language," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1708566, 2025, doi: 10.3389/frai. 2025.1708566.

[5] G. Kessikbayeva and I. Cicekli, "Rule-based morphological analyzer of Kazakh language," in Proc. 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, MD, USA. Association for Computational Linguistics, 2014, pp. 46–54, doi: 10.3115/v1/W14-2806.

[6] G. Tolegen, A. Toleu, and R. Mussabayev, "A finite state transducer-based morphological analyzer for the Kazakh language," in 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey. IEEE, 2022, pp. 165–170, doi: 10.1109/UBMK55850.2022.9919445.

[7] R. Cotterell and G. Heigold, "Cross-lingual character-level neural morphological tagging," in Proc. 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark. Association for Computational Linguistics, 2017, pp. 748–759, doi: 10.18653/v1/D17-1078.

[8] F. Godin, K. Demuynck, J. Dambre, W. De Neve, and T. Demeester, "Explaining character-aware neural networks for word-level prediction: Do they discover linguistic rules?" in Proc. 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, 2018, pp. 3275–3284, doi: 10.18653/v1/D18-1365.

[9] P. Dasari, A. Chelpuri, N. Vuppala, M. Marreddy, P. Krishnamurthy, and R. Mamidi, "Transformer-based context-aware morphological analyzer for Telugu," in Proc. Third Workshop on Speech and Language Technologies for Dravidian Languages, Varna, Bulgaria. Shoumen, Bulgaria: INCOMA Ltd., 2023, pp. 25–32, doi: 10.26615/978-954-452-085-4_004.

[10] G. Abudouwaili, K. Abiderexiti, N. Yi, and A. Wumaier, "Joint learning model for low-resource agglutinative language morphological tagging," in Proc. 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, Toronto, Canada. Association for Computational Linguistics, 2023, pp. 27–37, doi: 10.18653/v1/2023. sigmorphon-1.4.

[11] F. Riemenschneider and K. Krahn, "Heidelberg-Boston @ SIGTYP 2024 shared task: Enhancing low-resource language analysis with character-aware hierarchical transformers," in Proc. 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, St. Julian's, Malta. Association for Computational Linguistics, 2024, pp. 131–141, doi: 10.18653/v1/2024. sigtyp-1.16.

[12] K. Oflazer and C. Guzey, "Spelling correction in agglutinative languages," in Fourth Conference on Applied Natural Language Processing, Stuttgart, Germany. Association for Computational Linguistics, 1994, pp. 194–195, doi: 10.3115/974358.974406.

[13] A. Sorokin, "Spelling correction for morphologically rich language: A case study of Russian," in Proc. 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, Spain. Association for Computational Linguistics, 2017, pp. 45–53, doi: 10.18653/v1/W17-1408.

[14] A. Rozovskaya and D. Roth, "Grammar error correction in morphologically rich languages: The case of Russian," Transactions of the Association for Computational Linguistics, vol. 7, pp. 1–17, 2019, doi: 10.1162/tacl_a_00251.

[15] A. Kara, F. Marouf Sofian, A. Bond, and G. Şahin, "GECTurk: Grammatical error correction and detection dataset for Turkish," in Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, Nusa Dua, Bali. Association for Computational Linguistics, 2023, pp. 278–290, doi: 10.18653/v1/2023.findings-ijcnlp.26.

[16] S. Najafi, B. Hauer, R. R. Riyadh, L. Yu, and G. Kondrak, "Combining neural and non-neural methods for low-resource morphological reinflection," in Proc. CoNLL-SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection, Brussels, Belgium. Association for Computational Linguistics, 2018, pp. 116–120, doi: 10.18653/v1/K18-3015.

[17] A. Anastasopoulos and G. Neubig, "Pushing the limits of low-resource morphological inflection," in Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Association for Computational Linguistics, 2019, pp. 984–996, doi: 10.18653/v1/D19-1091.

[18] T. Bergmanis, K. Kann, H. Schütze, and S. Goldwater, "Training data augmentation for low-resource morphological inflection," in Proc. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection, Vancouver, Canada. Association for Computational Linguistics, 2017, pp. 31–39, doi: 10.18653/v1/K17-2002.

[19] A. Toleu, G. Tolegen, and A. Makazhanov, "Character-aware neural morphological disambiguation," in Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada. Association for Computational Linguistics, 2017, pp. 666–671, doi: 10.18653/v1/P17-2105.

[20] G. Tolegen, A. Toleu, and R. Mussabayev, "Contrastive learning for morphological disambiguation using large language models in low-resource settings," Applied Sciences, vol. 14, no. 21, Art. no. 9992, 2024, doi: 10.3390/app14219992.

[21] J. Traub, T. J. Bungert, C. T. Lüth, M. Baumgartner, K. H. Maier-Hein, L. Maier-Hein, and P. F. Jaeger, "Overcoming common flaws in the evaluation of selective classification systems," in Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024.

[22] M.-C. de Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal Dependencies," Computational Linguistics, vol. 47, no. 2, pp. 255–308, 2021, doi: 10.1162/coli_a_00402.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Non-intrusive Human Activity Recognition using Imbalanced mm-Wave Doppler Radar Streams and Ensemble of Time-Series Modeling Strategies

Khadija Adam Nasir
Dept. of Computer Science
Swansea University
Swansea, United Kingdom
2011842@swansea.ac.uk
ORCID: 0009-0009-2432-0088

Alaa Humaidan
Dept. of Computer Science
Swansea University
Swansea, United Kingdom
2030685@swansea.ac.uk
ORCID: 0009-0007-2753-5522

Sara Sharifzadeh
Dept. of Computer Science
Swansea University
Swansea, United Kingdom
Sara.Sharifzadeh@Swansea.ac.uk
ORCID: 0000-0003-4621-2917

Jeny Roy
Dept. of Computer Science
Swansea University
Swansea, United Kingdom
jenymaryroy@gmail.com

Abstract
Human Activity Recognition (HAR) enables automatic monitoring of human activities, with applications in intelligent healthcare, fall detection, and independent living for older people. Human activities occur in continuous sequences with seamless transitions; however, despite significant advances in radar-based HAR, continuous HAR remains underexplored, as many existing studies focus on single-activity windows rather than continuous activities. This paper addresses radar-based continuous HAR. We explored the efficiency of three time-series analysis strategies for continuous radar-based HAR, including 1) CNN-LSTM, 2) CNN-Transformer, and 3) CNN-Vision Transformer. We devised a fused ensemble learning approach that leverages the strengths of each pipeline by incorporating prior knowledge of per-class performance from each pipeline's evaluation. Our proposed ensemble method achieved the highest overall accuracy (84.81%) and better generalization, followed by the CNN-Vision Transformer and the CNN-LSTM. The natural imbalance condition of HAR data leads to prediction bias. In this paper, we address this problem and devise a data balancing and augmentation strategy based on averaging similar data windows.

Keywords: Human Activity Recognition (HAR), Doppler radar, continuous HAR, data balancing, Averaging, Ensemble, CNN-LSTM, CNN-Vision Transformer, CNN-Transformer

I. INTRODUCTION

The decline of cognitive and physical activity makes it difficult for some demographics, such as older people, to perform daily routine activities and increases their risk of accidents. According to the WHO, the population of people aged 60 and above will increase by nearly 50% [1]. Human activity recognition (HAR) provides a means to automatically explore, track, and categorize people's activities and actions. Both intrusive and non-intrusive methods can achieve HAR. Intrusive devices, such as digital cameras, invade users' privacy, while other devices, such as wearables, require users to wear them at all times. This makes them less than ideal devices to use, especially for elderly users with declining cognitive function who require independent living without compromising privacy. Therefore, the employment of non-intrusive devices is a far better idea. In recent years, digital sensors such as mm-wave Doppler radar have become a better alternative for continuous monitoring of



human activity while eliminating privacy invasion [2]. Another advantage of radar sensors is that they can operate under different lighting conditions without any loss of accuracy. For HAR, micro-doppler signatures comprising the movement of the head, torso, and limbs are extracted and analyzed [3]. One important, related challenge for HAR is that human activity data is inherently imbalanced. The huge imbalance in how human activities are naturally distributed might cause a model to be biased, favoring the most frequent activities and underperforming on rare ones. This can lead to very high overall accuracy despite failing to identify rare activities correctly. A fall is considered a rare activity, whereas standing or walking are very frequent. Missing a fall detection can come with dire and undesirable harm, especially for the elderly population. Furthermore, some activities might take longer, such as prolonged sitting or walking, while others might take less time, such as the transition from sitting to standing. Therefore, there is a need for balanced data to mitigate bias and improve recognition of underrepresented minority activities. Deep learning techniques such as convolutional neural networks (CNNs) have been used in HAR to extract features effectively from sensor data [4], and recurrent neural networks like Long Short-Term Memory (LSTM) are also promising for accuracy and performance on sequential data. This is due to their ability to learn and remember long-term dependencies in sequential data [5]. Other approaches combined CNNs and LSTMs, with the CNN layers extracting features for the LSTM layers. In wearable motion sensors [6], the hybrid CNN-LSTM approach achieved higher accuracy than using CNN or LSTM individually. In [7], a probabilistic neural network on single-subject activity data obtained from sensors was used to recognize activities. The limitation of these studies is that they utilized only single-activity recordings within a window. However, in real-life scenarios, activities occur continuously, one after another. Therefore, it would be valuable to extend these approaches to continuous HAR. Another promising approach for processing sequential data is vision transformers (ViTs), which modify the transformer architecture for image classification by treating image patches as sequential tokens [8]. This technique can be relevant to radar-based continuous HAR, where the data windows can be treated as sequences of image-like frames.

Most radar-based HAR studies utilized a single activity recording within a sample, e.g., sequences of frames, which are sometimes referred to as a window [3], [9]. On the other hand, some studies collect recordings from a series of various continuous activities. To investigate radar-based continuous HAR, the authors in [10] adopted two strategies for continuous activity recognition of radar data, using the Dynamic Time Warping (DTW) algorithm and LSTM. A hybrid CNN-LSTM approach was used by [11] for acoustic sensors, achieving good accuracy under non-noisy conditions. While some research has reported the adoption of some time-series analyses for continuous radar-based HAR, no clear comparison of such strategies has been conducted. Then, it is desirable to extend and perform a comparative study on these deep learning architectures that have demonstrated great potential.

Ensemble learning is the art of integrating multiple weak individual models to produce a model with improved performance and better generalization. Ensemble learning strategies such as unweighted averaging and majority voting are biased towards the stronger model's outcome [12]. Our proposed models have different strengths, with each model performing better on certain activity classes. This prior knowledge of per-class performance obtained during model evaluation is used to define class-specific experts among the models. Therefore, we perform ensemble learning based on how reliably each model predicts each class; we use class-specific performance as prior knowledge to guide the ensemble during inference.

This paper explores the efficiency of different time-series analysis strategies, including CNN-LSTM, CNN-Transformer, and CNN-Vision Transformer, for continuous radar-based HAR.

In this paper, we do: 1) highlight the natural imbalance condition in radar data and perform data augmentation by averaging, and data balancing with weighted cross-entropy loss during training to address the class-balance

challenge that drastically introduces prediction bias, 2) a comparison of several pipelines and devise a hybrid pipeline for Doppler radar-based continuous HAR for sequence to sequence modelling, and evaluate their overall performance and per class performance, 3) a fused ensemble learning approach to benefit from the strengths of each pipeline, by using prior knowledge of per-class performance obtained during each pipeline evaluation.

The dataset utilized for this research was collected at the Swansea University Living Lab. It contains radar recordings of different sequences of daily human activities.

II. METHODOLOGY

1. Data Description

The dataset used in this study is based on data collected using Doppler radar in the living lab at Swansea University, with 8 subjects aged 65 years and over; each subject conducted a series of activities 8 times. The data were then divided into overlapping windows (50% overlap), each with 15 timesteps and manually labeled per timestep. Some windows contain homogeneous activities, while others contain heterogeneous ones. A total of 1428 windows were generated. These were then divided into training (70%), validation (15%), and testing (15%) subsets. The split was based on random selection rather than a subject-wise basis. For this research, some activities from the experiment were omitted, and only five were used: sitting, standing, walking, falling, and bending. Figure 1 shows the sequence of activities.

2. Data Augmentation and Balancing

We utilized an averaging strategy for data augmentation to mitigate class imbalance and obtain more samples for the data-hungry pipeline (ViT). In the training dataset, for each class, a new window is created by averaging the features of every unique pair of homogeneous windows in that class. If the windows are similar, averaging occurs only when the pair has reached a 90% similarity threshold. Furthermore, generation is stopped when the total number of homogeneous windows for a particular class reaches 500. However, for class 3 (falling), the total number of homogeneous windows was set to 200. This is because the number of samples in that class is very limited, and repeatedly applying averaging can distort the class signals, causing much more severe confusion with other classes. Figure 2 shows the distribution of the training dataset before and after augmentation. Figure 3 shows a visualization of two windows and an averaged generated window.

3. Data Reshaping

The input data was originally shaped (15, 4096). Since all proposed pipelines require a CNN block, the input data was later reshaped to (15, 16, 256, 1) to make it compatible with conv2D layers in a convolutional neural network (CNN). Convolutional layers require a 2D input with channels to extract meaningful features.

4. Pipelines

The CNN-LSTM pipeline contains one CNN block applied at each timestep, followed by an LSTM layer that also outputs sequences of its hidden states at each timestep. The whole process is outlined in Figure 4.
CNN-Transformer The CNN-Transformer pipeline contains one CNN block applied per timestep, and the output is flattened into a 1D vector. This is followed by key components of the transformer: positional embedding, multihead attention, and a feed-forward network. The output features from the CNN block are linearly projected, and positional embeddings are added to capture temporal features. This is followed by two

transformer encoder layers applied sequentially. Each encoder contains the key elements of the transformer (multihead attention and feed-forward network).

CNN-Vision Transformer

This pipeline contains two CNN layers, applied per timestep, to extract features and reduce the spatial dimension, preparing the input for patch extraction. As required by vision transformers, features are converted into sequential tokens (patches), and positional embeddings are then applied. This is followed by four identical Vision Transformer encoder blocks, each containing multihead attention and a feed-forward network. Further, global average pooling is applied over the patch dimension at each timestep, and an output layer is applied to obtain sequence-to-sequence classification. The model architecture is shown in Figure 5. This model achieved the highest overall accuracy among the three pipelines.

The three pipelines above each have different strengths in prediction. To leverage all models, an ensemble approach is utilized. Since CNN-LSTM and CNN-vision transformer showed a better per-class accuracy, the two are chosen for ensemble learning. Each of the selected models has different strengths in predicting particular classes; therefore, a class-specific weighted fusion is used. This method outperforms the standard ensemble that assigns global weights regardless of the models' class accuracies. In our case, for example, CNN-LSTM achieves higher accuracy for the falling class, while CNN-Vision Transformer achieves higher accuracy for the standing, walking, and bending classes. Therefore, instead of assigning one weight per model, we first compute a performance matrix, which is the accuracy of each model for each class; we use this to obtain weights for each model per class. We define and compute class-specific w, c, m, son, alt, and simgeper models for each class (obtained by normalizing the performance matrix). The accuracy scores of all models are summed for each class, then each model's accuracy for that class is divided by this sum to ensure that the weights of a given class sum to 1 across all models.

Each model outputs a probability tensor, a $p_{i,t,c}^{(m)}$ is the probability assigned by model m for class c for each sample i at timestep t. We then generate a fused probability for each class, obtained by taking the weighted sum of the two models' predictions for that class. Since the weighted combination may not form a valid probability distribution, we further normalize the fused probability across all classes. The class with the highest probability is then considered the final prediction. Figure 6 shows the working mechanism of the ensemble learning model.

$$\text{fused}_{i,t,c} = \sum_{m=1}^{M} w_{c,m}\ p_{i,t,c}^{(m)} \tag{1}$$

normalise;

$$\hat{p}_{i,t,c} = \frac{\text{fused}_{i,t,c}}{\sum_{j=1}^{C} \text{fused}_{i,t,j}} \tag{2}$$

Finally,

$$y_{i,t} = \arg\max_{c} \hat{p}_{i,t,c} \tag{3}$$

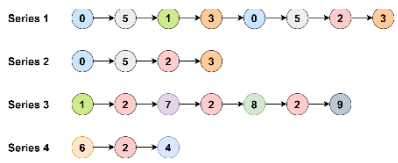


Figure 1. Example Original Series of activity sequences. The labels are as follows: 0-sitting, 1-standing, 2-walking, 3-falling, 4-bending, 5-sit to Stand, 6-standing and cooking, 7-standing and eating, 8-standing and opening/closing the microwave, 9-hitting the door.

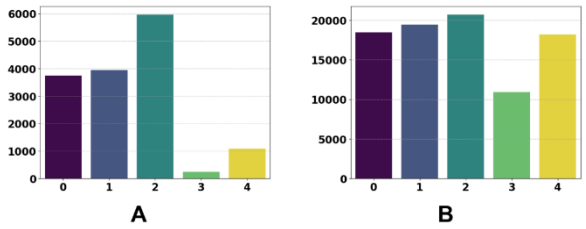


Figure 2. Distribution of the training dataset before and after augmentation. The X-axis represents the labels, and the Y-axis represents the number of occurrences. A represents the distribution of original data without augmentation, and B represents the distribution of data after augmentation. The labels are 0-sitting, 1-standing, 2-walking, 3-falling, 4-bending

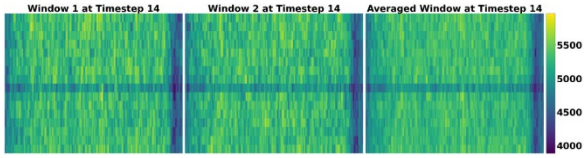


Figure 3. Comparison of two original windows with an averaged window. The averaged window is generated by averaging two similar windows of the same activity at each timestep.

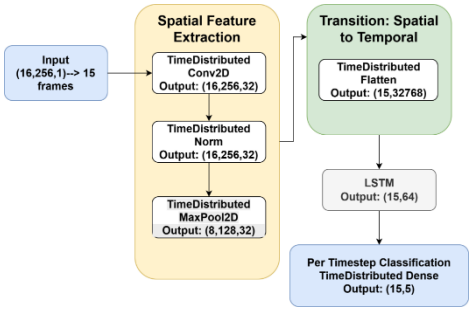


Figure 4. Processing workflow of CNN-LSTM Architecture. Convolutional block for feature extraction followed by an LSTM block and a per-timestep classification layer.

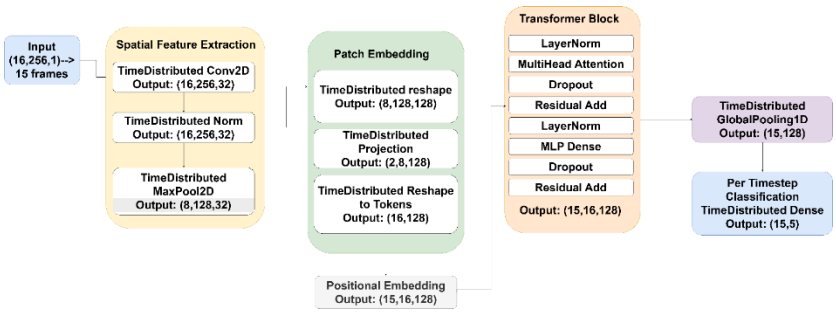


Figure 5. Processing Workflow of CNN-Vision Transformer architecture. Convolutional block for feature extraction followed by vision transformer blocks. The final block is a per-timestep classification layer.

III. RESULTS AND DISCUSSION

1. Results

In the first experiment, we evaluated the three pipelines on clean original data with no augmentation. To address class imbalance and improve the accuracy of the minority class, a weighted cross-entropy loss was used during training. In a preliminary experiment, we noticed that falling is consistently misclassified as walking. Since a fall activity usually follows a walk activity, confusion between the two classes persists. Hence, the weight of the minority fall class was manually set to 7, approximately 14 times that of the majority class (Walking). Falling cases are critical, and misclassifying them carries a higher risk than misclassifying other classes.

CNN-LSTM achieved an accuracy of 82.42%, CNN-Transformer achieved 49.22%, and CNN-Vision Transformer achieved an accuracy of 71.65%. The overall accuracy of the CNN-LSTM was higher under the unaugmented data condition. Moreover, CNN-Transformer had 0 detections of the minority falling class. The accuracy of CNN-Transformer and CNN-Vision Transformer is low because both require large amounts of data to perform at their full potential. The per-class accuracy results for this experiment are summarised in Table 1.

In the second experiment, we repeated the first experiment using the augmented train set defined in the methodology section. The number of data windows increased to 5845. Since the minority-falling class samples have now increased, we adjusted the class weights to approximately 9 times those of the majority class. CNN-LSTM achieved a slightly lower accuracy of 82.20%; CNN-Transformer increased by close to 35% to 83.68%, while CNN-Vision Transformer increased to 84.34%. However, despite the major increase in CNN-Transformer, its failure to accurately classify the minority class persists. The per-class accuracy results for this experiment are summarised in Table 2.

In the third experiment, we selected the two best models, specifically in terms of class accuracy (the non-augmented CNN-LSTM and the augmented CNN-Vision Transformer), for ensemble learning. We leveraged the best qualities of each model to guide ensemble predictions. This experiment achieved the highest overall accuracy, at 84.81%. The results of this experiment are summarised in Figure 7. Table 3 compares the performance of the two selected models and the ensemble model for minority-class detection (falling).

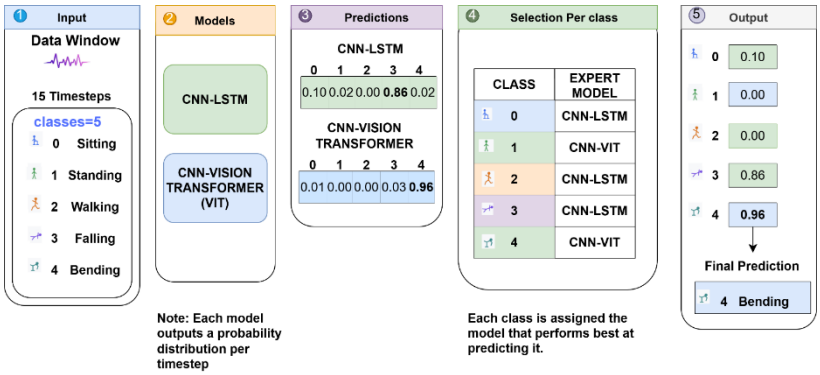


Figure 6. Ensemble model working mechanism: class probability vectors are obtained from candidate models for each timestep classification. An expert model for each class is assigned based on prior knowledge obtained from the performance matrix. The probability for each class is extracted from its assigned expert and combined into a new probability vector used for the final prediction. The class with the highest probability is the final prediction.

2. Discussion

The three different experiments reveal an interesting pattern. CNN-LSTM achieves the highest accuracy when only the weighted cross-entropy loss is used, and no augmentation is applied. This aligns with the findings in [11] for acoustic HAR. Although it achieved the highest accuracy for the falling class (our minority class), the excessive number of false alarms led to a precision of only 23%, resulting in the lowest F1-score at 0.38 and the lowest overall accuracy. The excessive number of false alarms could lead users to mistrust the model in real-life situations. In contrast, the performance of CNN-Vision Transformer increased after augmentation. This indicates that the proposed augmentation strategy is effective and that additional training data improved classification performance across all classes. Furthermore, the precision of the falling class increased to 49%, indicating fewer false-positive alarms, and the macro F1 score increased to 0.81, indicating balanced performance across all classes. A similar trend of performance improvement was also demonstrated in CNN-Transformer after augmentation. On the other hand, the ensemble model achieved the highest overall accuracy at 84.81% and had the strongest balance among all classes, with a macro F1 of 0.811, a slight increase over CNN-Vision Transformer, and a weighted F1 score of 0.85. Although CNN-LSTM achieves the highest fall-class detection, the ensemble model results indicate that a modest reduction in fall detection yields a middle ground with slightly lower accuracy but better generalization and a more balanced HAR.

Table 1. Per-class Accuracy on Unaugmented Data

Activity/Class	CNN-LSTM	CNN-Transformer	CNN-Vision Transformer
Sitting	**97.96**	25.61	79.49
Standing	**77.18**	54.87	93.86
Walking	**73.45**	48.46	47.09
Falling	**96.55**	0.00	77.78
Bending	**94.07**	81.36	95.34

Table 2. Per-class Accuracy on Augmented Data

Activity/Class	CNN-LSTM	CNN-Transformer	CNN-Vision Transformer
Sitting	**97.48**	96.39	93.87
Standing	87.28	93.27	**89.65**
Walking	70.14	69.31	**73.15**
Falling	34.48	17.24	**82.76**
Bending	84.75	85.17	**95.76**

Table 3. Comparison of Models in Terms of Minority Class Detection

Metric	CNN-LSTM	CNN-Transformer	CNN-Vision Transformer
Overall accuracy	82.42%	84.34%	**84.81%**
Weighted F1-Score	0.8300	0.8400	**0.8485**
Macro F1-score	0.7500	0.8100	**0.8107**
Precision	0.2300	0.4900	0.4545
Recall	**0.9655**	0.8276	0.8621
F1-score	0.3800	**0.6200**	0.5952

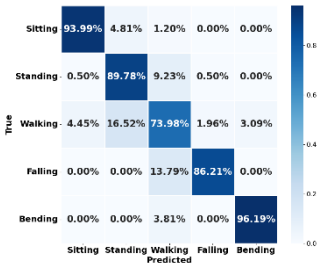


Figure 7. Confusion matrix for the Ensemble Model

IV. CONCLUSION

This study explored radar-based HAR in a continuous setting. 1) We addressed the natural imbalance condition of HAR data and devised a data balancing and augmentation strategy by averaging single activity windows that are at least 90% similar at every timestep, 2) We evaluated and compared three pipelines: CNN-LSTM, CNN-Transformer, and CNN-Vision Transformer, 3) We developed a fused ensemble learning strategy that leverages

the strengths of each pipeline by using prior knowledge of per-class performance obtained during each pipeline's evaluation. On non-augmented data, CNN-LSTM achieved remarkable accuracy using only a weighted cross-entropy loss. In the augmented data, we observed a significant increase in the accuracy of the CNN-Transformer and CNN-Vision Transformer models. This indicated that the augmentation and amount of data influence these architectures. The CNN-Vision Transformer achieves higher accuracy and better generalization across classes. In both augmented and non-augmented data, CNN-Transformer achieves lower accuracy, particularly due to its failure to detect and predict the minority class. Our proposed ensemble method achieved the highest overall accuracy, with a slight drop in accuracy for the fall class; however, this yields better generalization and a more trusted, reliable, and balanced HAR. For future work, we would like to extend this work to adapt it to a different radar environment with different subjects and different age groups.

REFERENCES

[1] 'Aging and health'. Accessed: Jul. 23, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[2] Z. Li, J. Le Kernec, Q. Abbasi, F. Fioranelli, S. Yang, and O. Romain, 'Radar-based human activity recognition with adaptive thresholding towards resource-constrained platforms', Sci. Rep., vol. 13, no. 1, p. 3473, 2023, doi: 10.1038/s41598-023-30631-x.

[3] H. Li, A. Shrestha, H. Heidari, J. L. Kernec, and F. Fioranelli, 'Activities Recognition and fall detection in continuous data streams using radar sensor', in 2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC), 2019, pp. 1–4. doi: 10.1109/IMBIOC.2019.8777855.

[4] A. Bevilacqua, K. MacDonald, A. Rangarej, V. Widjaya, B. Caulfield, and T. Kechadi, 'Human Activity recognition with convolutional neural networks', Jun. 2019, doi: 10.1007/978-3-030-10997-4_33.

[5] D. Singh, E. Merdivan, J. Kropf, and A. Holzinger, 'Class imbalance in multi-resident activity recognition: an evaluative study on explainability of deep learning approaches', Univers. Access Inf. Soc., 2024, doi: 10.1007/s10209-024-01123-0.

[6] E. Koşar and B. Barshan, 'A new CNN-LSTM architecture for activity recognition employing wearable motion sensor data: Enabling diverse feature extraction', Eng. Appl. Artif. Intell., vol. 124, Sep. 2023, doi: 10.1016/j.engappai.2023.106529.

[7] L. G. Fahad and S. F. Tahir, 'Activity recognition and anomaly detection in smart homes', Neurocomputing, vol. 423, pp. 362–372, Jan. 2021, doi: 10.1016/j.neucom.2020.10.102.

[8] A. Dosovitskiy et al., 'An Image is worth 16x16 words: Transformers for image recognition at scale', ArXiv, vol. abs/2010.11929, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:225039882

[9] Y. Kim, I. Alnujaim, and D. Oh, 'Human activity classification based on point clouds measured by millimeter wave MIMO radar with deep recurrent neural networks', IEEE Sens. J., vol. 21, no. 12, pp. 13522–13529, 2021, doi: 10.1109/JSEN.2021.3068388.

[10] R. Mehta, S. Sharifzadeh, V. Palade, B. Tan, A. Daneshkhah, and Y. Karayaneva, 'Deep learning techniques for radar-based continuous human activity recognition', Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1493–1518, 2023, doi: 10.3390/make5040075.

[11] A. Humaidan, J. Roy, S. Sharifzadeh, R. Mehta, A. Tales, and J. Macinnes, 'Exploring human activity recognition with acoustic data: a comparative study of CNN-LSTM, ViViT, and ResNet-temporal transformer model', in 2025 IEEE Symposium on Computational Intelligence in Image, Signal Processing and Synthetic Media Companion (CISM Companion), 2025, pp. 1–5. doi: 10.1109/CISMCompanion65074.2025.11032695.

[12] M. A. Ganaie, M. Hu, M. Tanveer, and P. N. Suganthan, 'Ensemble deep learning: A review', CoRR, vol. abs/2104.02395, 2021, [Online]. Available: httpsw2qt://arxiv.org/abs/2104.02395

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Full-Frame Sparse Temporal Multitask Learning for Student Affective State Recognition Using DAiSEE

İbrahim DOLAŞ
Gazi University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering, Ankara, Türkiye
dolasibrahim382@gmail.com
ORCID ID: 0009-0000-8156-5369

Hüseyin POLAT
Gazi University, Department of Computer Engineering, Ankara, Türkiye
polath@gazi.edu.tr
ORCID ID: 0000-0003-4128-2625

Abstract

Automatic analysis of learner affect can support online learning by providing complementary information about behavioral and emotional states that are difficult for instructors to observe continuously. This study presents a reproducible, full-frame, sparse-temporal, multitask framework for student affective-state recognition using the DAiSEE dataset. The official Train, Validation, and Test partitions are preserved, and each usable clip is represented by 16 temporally distributed RGB frames, resized to 160 × 160 pixels, and associated with 4 clip-level targets: Engagement, Boredom, Confusion, and Frustration. The completed audit retained 5,357 training, 1,429 validation, and 1,784 test clips, yielding 8,570 sequences and 137,120 images. The proposed framework uses a shared MobileNetV2-GRU representation with four task-specific classification heads, while a single-frame baseline, temporal configurations, and imbalance-aware training are defined for controlled comparison. The dataset analysis reveals substantial, target-specific class imbalance, demonstrating that conventional accuracy alone can yield a misleading assessment of affective-state recognition. Therefore, the evaluation protocol emphasizes balanced and ordinal measures together with class-wise analysis. The reported quantitative results are limited to the verified majority-class validation baseline, which establishes a reference point for evaluating the trained models. The study provides an audited data pipeline, transparent experimental protocol, and reproducible foundation for subsequent temporal and multimodal student-state analysis.

Keywords: DAiSEE, student affective state, multitask learning, temporal modeling

I. INTRODUCTION

Automatic engagement estimation has become increasingly relevant as learning activities move between physical and digital environments. In a face-to-face classroom, instructors can combine eye contact, posture, note-taking, response timing, and interaction history when judging whether a learner is following the lesson. Most of these cues are difficult to monitor continuously in a large online class. A systematic review of automatic engagement estimation, therefore, identifies engagement recognition as a potential component of adaptive and personalized learning support, while emphasizing that engagement is multidimensional and must be interpreted through observable affective and behavioral cues [1]. DAiSEE was introduced as an in-the-wild multi-label video dataset containing 9,068 clips from 112 users. Each clip is annotated for Engagement, Boredom, Confusion, and Frustration at four intensity levels: very low, low, high, and very high [2]. This formulation is richer than binary engaged-versus-disengaged recognition because it represents several related learner states and preserves intensity. It is also more difficult: the labels are assigned at the clip level, visible expressions may be subtle, backgrounds and illumination vary, and neighboring ordinal levels can be visually similar. A central methodological challenge is the strong class imbalance in the official labels. Prior DAiSEE work has shown that dominant classes can obscure weak minority-class performance and motivate balancing

strategies [3]. A second challenge is ordinality: predicting level 2 rather than level 3 should not be treated as severely as predicting level 0 rather than level 3. Ordinal engagement research consequently supports reporting measures such as quadratic weighted kappa and mean absolute error in addition to exact classification accuracy [4]. These issues make a majority-class reference and macro-averaged metrics essential for credible evaluation. This paper develops a full-frame sparse temporal multitask pipeline and reports on the parts that generated files have already verified. The main contributions are fourfold: first, a validated preprocessing stage that links sixteen ordered full-frame images to the four original labels; second, a shared MobileNetV2-GRU design that provides a reproducible basis for static-versus-temporal and weighted-versus-unweighted comparisons; third, a transparent majority-class baseline showing how ordinary accuracy can hide minority-class failure; and fourth, a clearly defined multimodal journal extension based on calibrated facial, iris, head-motion, and posture cues. The learned-model protocol is specified without presenting numerical claims that the experiment pipeline has not yet produced.

II. RELATED WORK

1. Video-Based and Temporal Engagement Recognition

Engagement varies over time, so video-based recognition is often formulated as a spatiotemporal learning problem. Abedi and Khan combined ResNet frame features with a temporal convolutional network and showed that temporal variation across frames can improve four-level DAiSEE engagement recognition [5]. This design principle is relevant to the present work because a single image may capture a temporary expression or head position. In contrast, an ordered sequence can represent persistence, transitions, and small behavioral changes across the clip. Other studies have explored more computationally intensive combinations of spatial and temporal learning. Mehta et al. used three-dimensional DenseNet and self-attention to model engagement-related information and reported experiments for all four DAiSEE targets [6]. In contrast, Bag of States questioned whether the exact temporal order is always necessary and showed that non-sequential state occurrence can also be informative [7]. This disagreement motivates the static-versus-temporal comparison in the present protocol rather than assuming that a recurrent model must automatically be superior.

2. Multitask and Shared Representation Learning

The four DAiSEE targets describe the same video clip and may share visual evidence. For example, reduced engagement can co-occur with boredom, while confusion and frustration may appear together during a difficult task. Multitask engagement research has used auxiliary objectives to regularize representation learning and reduce overfitting [8]. DenseAttNet also examined the four DAiSEE labels within a common modeling context [6]. Click or tap here to enter text. These studies support investigating whether a shared feature extractor can learn information that is useful across related affective states. The multitask formulation used here is deliberately simple. It does not reproduce the triplet-loss design of [8]; instead, the same MobileNetV2-GRU representation feeds four independent categorical heads. This distinction is important because multitask learning is beneficial only when shared features transfer across tasks without causing negative transfer. For that reason, the framework saves per-target metrics as well as averages, so a high mean score cannot conceal degradation in one of the four outputs.

3. Imbalance-Aware and Ordinal Learning

Class imbalance is a major risk in DAiSEE. Santoni et al. explicitly examined imbalanced engagement labels and employed feature-level processing and resampling to reduce majority-class bias [3]. Synthetic

oversampling is not directly applied to the RGB images in the present pipeline because generated image samples could introduce unrealistic visual content. Instead, the training code supports class weighting at the loss level. The principal weighting option follows the effective-number concept proposed by Cui et al., in which the contribution of additional samples diminishes as the class size increases [9]. The resulting weights are less extreme than raw inverse-frequency weights and can therefore reduce unstable gradients. Because the labels are ordered, the evaluation combines balanced classification measures with QWK and MAE [4]. Together, these choices address two different problems: imbalance determines how strongly each class contributes during learning, whereas ordinal metrics determine how prediction distance is interpreted during evaluation.

III. DATASET AND PREPROCESSING

The project uses the official DAiSEE Train, Validation, and Test label files. The original clips are approximately ten seconds long [2]. Each raw video is matched to its ClipID, and its four labels are read from the corresponding CSV row. Although the physical image folders are organized by Engagement for storage convenience, the unified metadata file provides Engagement, Boredom, Confusion, and Frustration simultaneously. This avoids duplicating the same frame sequence into four separate label structures. Sixteen frame positions are distributed across each usable clip to form a sparse temporal representation. Every selected frame is converted to RGB, resized with aspect-ratio-preserving padding to 160 × 160 pixels, and stored as a JPEG image. The study uses only the full-frame webcam view; no upper-body crop, facial-only crop, or pose-landmark representation is included. Full-frame input retains the learner, visible body region, desk area, and environmental context, but it can also expose the model to background correlations. This trade-off is considered in the limitations.

Table 1. Audited full-frame sequence dataset

Split	Usable clips	Frames per clip	Image size
Train	5,357	16	160 × 160 × 3
Validation	1,429	16	160 × 160 × 3
Test	1,784	16	160 × 160 × 3
Total	8,570	16	160 × 160 × 3

Table 1 summarizes the completed audit. The extraction and integrity checks retained 5,357 Train, 1,429 Validation, and 1,784 Test sequences, for a total of 8,570 complete clips and 137,120 stored images. Figure 1 illustrates one ordered sequence. The learner remains the same throughout the sixteen frames, while small changes in head position, gaze direction, hand location, and posture provide the temporal information that a recurrent model may learn. The repeated appearance of the same person is therefore expected and does not indicate duplicate samples.

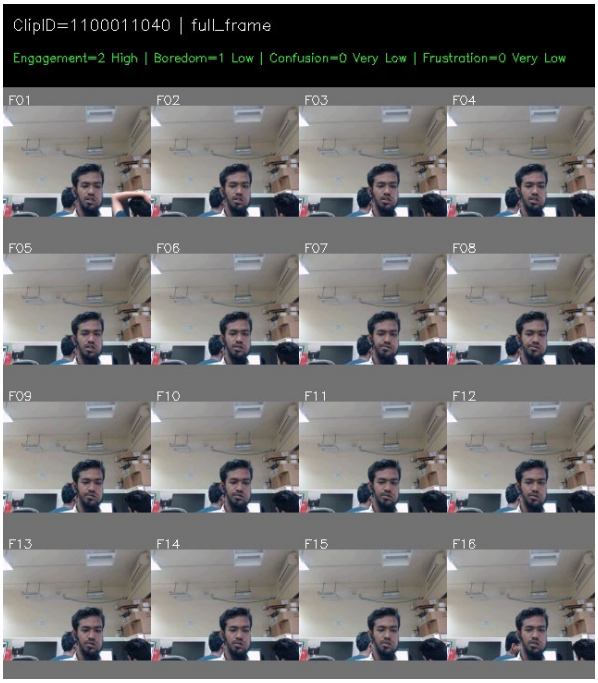


Figure 1. Example ordered sixteen-frame full-frame sequence extracted from one DAiSEE clip.

Table 2. Training-set class distribution for the four affective-state targets

Target	Level 0	Level 1	Level 2	Level 3
Engagement	34	213	2,617	2,493
Boredom	2,433	1,696	1,072	156
Confusion	3,616	1,245	430	66
Frustration	4,182	941	191	43

The distributions in Table 2 and Figure 2 are strongly skewed. Engagement levels 2 and 3 account for 95.39% of the training clips, while levels 0 and 1 together account for only 4.61%. Boredom is less concentrated but still contains only 156 level-3 examples. Confusion level 0 represents 67.50% of the training data, and Frustration level 0 represents 78.07%, whereas their level-3 classes contain only 66 and 43 clips, respectively. These differences explain why a model can obtain apparently high accuracy by repeatedly predicting a dominant class. The imbalance pattern is also target-specific. High-intensity labels dominate engagement, whereas very-low labels dominate Confusion and Frustration. Consequently, a single global balancing decision may not affect all tasks equally. The multitask loader, therefore, computes weights separately for each output head. The Validation and Test partitions are never used to calculate training weights, preserving the separation between model fitting and evaluation.

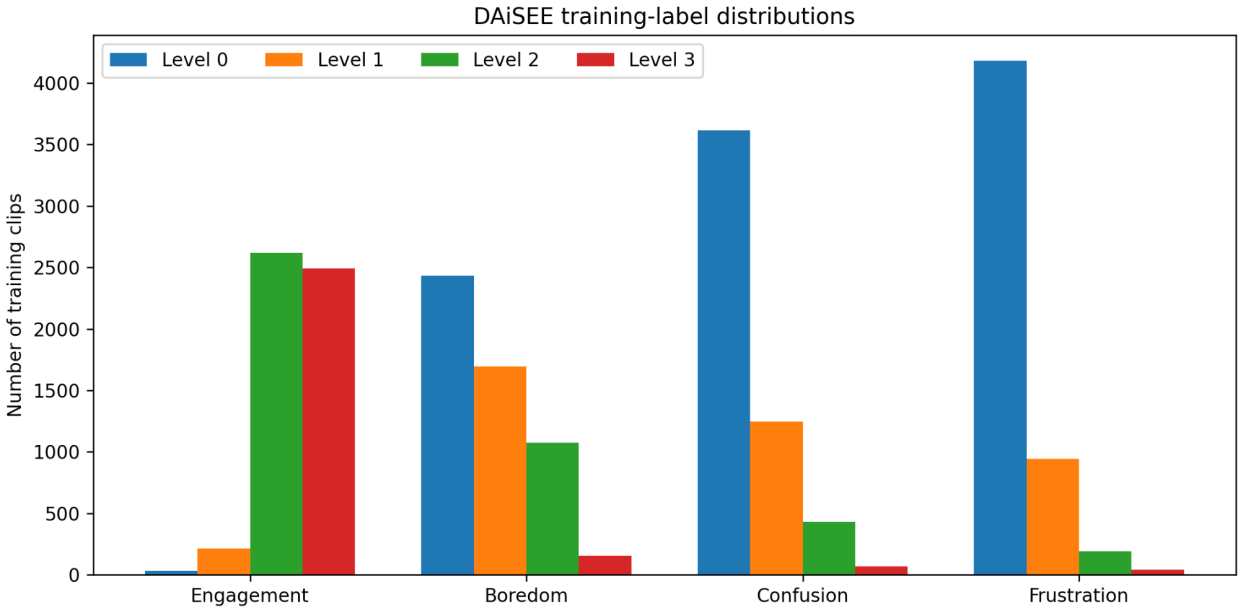


Figure 2. Training-label distributions for Engagement, Boredom, Confusion, and Frustration.

IV. PROPOSED FRAMEWORK

1. Sparse Temporal Input and Frame Encoder

For each clip, the model input is an ordered tensor X = {x1, x2, ..., xT}, where each xt is a 160 × 160 RGB image. Sixteen images are retained in storage, but the training loader can select T = 4, 8, or 16 evenly distributed positions. The primary conference configuration uses T = 8 to balance temporal coverage and computational cost. Each frame is normalized using the selected backbone and then passed through the same TimeDistributed MobileNetV2 encoder. MobileNetV2 is selected because its inverted residual and linear bottleneck designs reduce memory and computational cost for resource-constrained inference [10]. The frame encoder produces one feature vector per sampled image. Sharing the same encoder across time ensures that frame-level representations are comparable and avoids multiplying the number of trainable backbones by the sequence length.

2. Temporal and Multitask Prediction

The final recurrent state is passed through a shared dense representation, followed by four independent softmax layers. Each output contains four probabilities corresponding to levels 0-3 for Engagement, Boredom, Confusion, or Frustration. Task-specific heads allow the shared temporal features to be reused while preserving different decision boundaries for each label. The total multitasking objective is the sum of the four task losses. The ordinary baseline uses sparse categorical cross-entropy without weighting. The imbalance-aware configuration applies effective-number class weights separately to each target [9]. During transfer learning, the MobileNetV2 backbone is first trained with its convolutional layers frozen and can subsequently be fine-tuned at a smaller learning rate. Early stopping, learning-rate reduction, and best-checkpoint saving are used to limit overfitting and preserve the strongest Validation checkpoint.

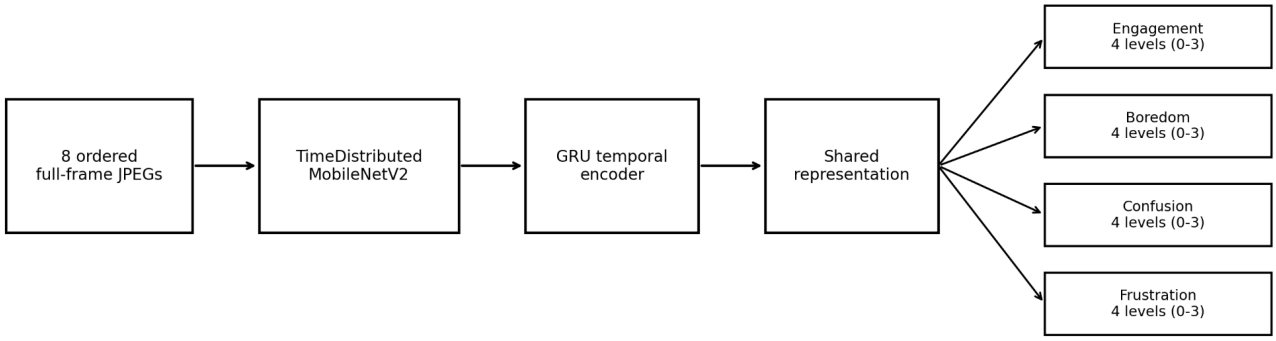


Figure 3. Proposed full-frame sparse temporal multitask architecture.

A static MobileNetV2 model using one central frame is retained as the spatial baseline. The temporal comparison is necessary because a GRU introduces additional parameters and may learn little when consecutive frames are nearly identical. Conversely, subtle transitions that are invisible in one frame may improve the temporal model. The static and temporal configurations, therefore, use the same full-frame preprocessing and output heads so that the principal architectural difference is temporal aggregation. The architecture is also designed for later connection to a rolling webcam buffer. A deployment implementation must reuse the training image size, channel order, normalization, frame-selection rule, and temporal window. Webcam output should initially be interpreted as a qualitative feasibility demonstration, as a live prediction is not quantitative evidence of accuracy without independently labeled participants and a predefined evaluation protocol.

V. EXPERIMENTAL DESIGN AND REPRODUCIBILITY

The official Train split is reserved for optimization, Validation is used for configuration selection, and Test must remain untouched until the final configuration has been locked. Each image associated with a given ClipID remains in the same official split. The supplied experiment design contains four controlled comparisons: a one-frame static MobileNetV2 baseline, an eight-frame MobileNetV2-GRU model without weighting, the same temporal model with effective-number weighting, and a 4/8/16-frame ablation. These comparisons are defined in the reproducibility package, but this paper reports only metrics that have already been generated and verified. For each target, the protocol records accuracy, balanced accuracy, macro precision, macro recall, macro F1-score, quadratic weighted kappa (QWK), mean absolute error (MAE), within-one-level accuracy, classwise recall, and a confusion matrix. Macro F1 and balanced accuracy give rare classes equal importance, whereas QWK and MAE exploit the ordered nature of the four levels. A model should not be selected from accuracy alone, and no Test result should be used to tune frames, weighting, loss, or the temporal encoder. Every run is stored in an independent output directory containing the configuration, random seed, model summary, learning history, checkpoints, predictions, and evaluation files. This separation makes the analysis auditable and prevents Validation and Test outputs from being mixed. The present conference manuscript can therefore be updated directly from generated CSV, JSON, and PNG files while retaining the same definitions and split policy in a later extended study.

VI. DATASET AUDIT AND MAJORITY-CLASS BASELINE RESULTS

1. Dataset Integrity and No-Training Reference

The completed audit verifies all four label ranges, checks the official split counts, and links each usable ClipID to a complete sequence of sixteen full-frame images. The resulting metadata contains 8,570 clips and 137,120 image files. Figure 1 and Figure 2 provide qualitative and quantitative checks on the extracted representation: the sequence figure confirms temporal order within a single clip, while the combined distribution chart reveals that the direction of imbalance differs substantially across targets. The majority class is estimated separately from the Train split and then assigned to every Validation clip. Engagement, therefore, receives a constant level-2 prediction, whereas Boredom, Confusion, and Frustration receive constant level-0 predictions. This simple procedure is deliberately incapable of recognizing minority levels, but it provides an essential lower bound and exposes the extent to which the label distribution can inflate ordinary accuracy. Averaged across the four tasks, the baseline obtains 57.02% accuracy, 25.00% balanced accuracy, 17.79% macro F1, 0.0000 QWK, 0.6149 MAE, and 84.34% within-one-level accuracy. The high within-one-level score occurs because many errors lie near the dominant ordinal level; it must not be interpreted as reliable four-class recognition. Table 3 reports the per-target values used to calculate these averages. The per-target results show two different failure patterns. Engagement appears moderately accurate because levels 2 and 3 dominate, while Confusion and Frustration appear highly accurate because level 0 dominates. In every case, balanced accuracy remains exactly 25%, which is chance-level recall averaged over four classes for a constant predictor. The zero QWK values further confirm that the baseline does not preserve meaningful ordinal agreement.

Table 3. Majority-class baseline on the Validation split

Target	Majority output	Accuracy	Bal. Acc.	Macro F1	MAE	Within ±1
Engagement	High (2)	56.89%	25.00%	18.13%	0.4472	98.39%
Boredom	Very Low (0)	31.21%	25.00%	11.89%	1.2050	57.52%
Confusion	Very Low (0)	65.92%	25.00%	19.87%	0.4647	88.45%
Frustration	Very Low (0)	74.04%	25.00%	21.27%	0.3429	93.00%
Average	**-**	**57.02%**	**25.00%**	**17.79%**	**0.6149**	**84.34%**

The baseline is therefore useful as an evaluation safeguard rather than as a practical classifier. Any learned model should primarily outperform it in macro F1, balanced accuracy, QWK, and minority-class recall. An increase in accuracy without improvement in these measures would indicate that the model has learned the same majority bias rather than the affective distinctions required by the task.

VII. DISCUSSION

The verified baseline demonstrates the central evaluation problem in DAiSEE: a high ordinary accuracy can coexist with no useful recognition of minority levels. Constant prediction of very-low Frustration is correct for 74.04% of Validation clips, but balanced accuracy remains 25%, and QWK remains zero. The same issue appears for Confusion. Consequently, conclusions must be driven by macro F1, balanced accuracy, QWK, MAE, per-class recall, and confusion matrices rather than by a single accuracy value. The sparse temporal

representation offers an explicit efficiency-versus-detail trade-off. Sixteen stored frames cover the entire 10-second clip while reducing storage and preprocessing compared with dense video decoding. The eight-frame main configuration further limits training memory. This is appropriate for rapid experimentation and later webcam deployment, although brief micro-behaviors between sampled positions may be missed. The 4/8/16-frame ablation is therefore necessary before concluding that one temporal density is sufficient. Full-frame input preserves facial appearance, upper-body posture, hand movement, desk interaction, and contextual information. That context may improve recognition, but it also creates a risk of shortcut learning from identity, room, illumination, clothing, or recording background. A strong journal extension should therefore include subject-disjoint evaluation, background sensitivity checks, face-masked or context-masked ablations, and explanation maps that test whether predictions depend on the learner rather than the environment. The current study has four principal limitations. First, DAiSEE labels are assigned at the clip level, so individual frames may not visibly exhibit the annotated state. Second, the rarest classes contain very few training examples, making estimates of classwise performance uncertain. Third, the majority baseline is a lower bound and does not establish the effectiveness of the proposed learned architecture. Fourth, a webcam demonstration without independent labels is qualitative and should not be presented as proof of real-world accuracy. Despite these limitations, the completed audit, figures, metadata, and exact baseline metrics provide a defensible foundation. They define the data that enters the model, expose the imbalance that any method must overcome, and specify which metrics are required for a fair comparison. This evidence is sufficient for a transparent framework-and-baseline study, while learned-model and multimodal claims remain reserved for experiments that can be independently reproduced.

VIII. FUTURE WORK

The next stage of this research will extend the full-frame temporal framework with complementary multimodal behavioral cues extracted from a webcam. MediaPipe-based facial landmarks will be used to derive facial geometry, eye openness, relative iris movement, and head-pose dynamics, including yaw, pitch, and roll. A short participant-specific calibration procedure will be investigated to normalize relative iris displacement and reduce sensitivity to individual differences in camera position and natural gaze behavior. Upper-body and shoulder-pose features may also be incorporated when reliable pose landmarks are available. These modalities will be fused with the visual MobileNetV2-GRU representation using controlled ablation experiments. The extended study will compare full-frame, geometric, and fused configurations under subject-independent evaluation and will examine class-wise performance, ordinal metrics, calibration, failure cases, and computational cost. A separately collected webcam dataset with independently obtained ground-truth annotations will also be required to evaluate real-world generalization. These experiments will determine whether multimodal fusion improves recognition reliability; such improvement is treated as a hypothesis for future evaluation rather than as a result claimed in the present study. Privacy and responsible use will be considered throughout the extension. Whenever possible, processing will be performed locally, and raw webcam video will not be retained after feature extraction. The system will be designed as an assistive analysis tool rather than a mechanism for grading, surveillance, identity recognition, or definitive inference of a learner's internal state.

IX. CONCLUSION

This paper presented a reproducible, full-frame, sparse-temporal, multitask framework and an audited majority-class baseline for recognizing Engagement, Boredom, Confusion, and Frustration on the DAiSEE dataset. The completed preprocessing audit retained 5,357 training, 1,429 validation, and 1,784 test clips, resulting in 8,570 complete sequences and 137,120 stored images. The official dataset partitions were

preserved, and the preprocessing and evaluation procedures were defined to support reproducible temporal experiments. The verified majority-class baseline demonstrates the central difficulty of DAiSEE: conventional accuracy can appear relatively high under strong class imbalance while providing little evidence of minority-class recognition. Therefore, balanced, macro-averaged, ordinal, and class-wise measures are necessary for meaningful evaluation. The reported results establish a transparent lower bound against which the planned MobileNetV2-GRU experiments can be compared. The proposed framework provides a lightweight foundation for temporal multitask learning while avoiding unsupported claims about model performance. Future work will evaluate the trained temporal models and extend the framework to incorporate calibrated facial, relative iris, head-motion, and posture features via multimodal fusion. These extensions will be evaluated through controlled ablations and independently labeled data to determine whether additional behavioral modalities improve generalization and reliability.

REFERENCES

[1] S. N. Karimah and S. Hasegawa, "Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods," Dec. 01, 2022, Springer. doi: 10.1186/s40561-022-00212-y.

[2] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild," Jul. 2022, [Online]. Available: http://arxiv.org/abs/1609.01885

[3] M. M. Santoni, T. Basaruddin, and K. Junus, "Convolutional Neural Network Model-based Students' Engagement Detection in Imbalanced DAiSEE Dataset." [Online]. Available: www.ijacsa.thesai.org

[4] A. Abedi and S. Khan, "Affect-driven Ordinal Engagement Measurement from Video," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2106.10882

[5] A. Abedi and S. S. Khan, "Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network," in Proceedings - 2021 18th Conference on Robots and Vision, CRV 2021, Institute of Electrical and Electronics Engineers Inc., May 2021, pp. 151–157. doi: 10.1109/CRV52889.2021.00028.

[6] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of students' engagement," Applied Intelligence, vol. 52, no. 12, pp. 13803–13823, Sep. 2022, doi: 10.1007/s10489-022-03200-4.

[7] A. Abedi and S. S. Khan, "Engagement Measurement Based on Facial Landmarks and Spatial-Temporal Graph Convolutional Networks," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2403.17175

[8] O. Copur, M. Nakıp, S. Scardapane, and J. Slowack, "Engagement Detection with Multi-Task Training in E-Learning Environments," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.04020

[9] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.05555

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1801.04381

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Visual Editing and Debugging of Feature Models

Cheng Thao
Dept. of Computer Science & Cybersecurity
Metropolitan State University
St. Paul, MN 55106, USA
cheng.thao@metrostate.edu

Ethan V. Munson
Dept. of Electrical Engineering& Computer Science
University of Wisconsin-Milwaukee
Milwaukee, WI 53201-0784, USA
munson@uwm.edu

Abstract
In software product line engineering (SPLE), a feature model is used to model commonality and variability in a family of products. It is used in all phases of SPLE to define product line scope, requirements, and documentation; to configure products; and to generate product code. Despite the importance of feature models in supporting SPLE, tools for feature model editing and meaningful debugging are lacking, hindering SPLE adoption. We propose a visual feature model editor that provides non-intrusive visual debugging in terms of features and their relations.

Keywords*:* software product line engineering; feature model

I. INTRODUCTION

Software product line engineering (SPLE) is a methodology for developing a family of software products within a particular domain by systematically reusing shared code to improve product quality and reduce development time and cost [1], [2]. A feature is a characteristic, quality, or functionality of a product. Thus, it is more natural to describe a product in terms of its features than in terms of lower-level implementation and design. A feature model describes a product family in terms of common and variant features, as well as any relationships or constraints among them. A feature model should represent all possible products within the product family. Feature models are used in all phases of SPLE: to define product line scope, to create requirements and documentation, to drive or guide the product configuration process, and, in some cases, to generate product code. Because feature models are described using a formal language, automated analysis can be performed to determine the number of possible products or to detect model inconsistency, dead features, or false optional features [3]. A feature model is *inconsistent* when multiple constraints in the model are violated simultaneously in such a way that no product can be represented. When a model becomes inconsistent, the user must debug it to find and correct errors. Current tools for feature models offer little or no support for debugging. They provide little or no explanation of the causes of inconsistency or of any changes that could be made to resolve the inconsistency [3]. These tools can indicate that the model is inconsistent, but the user must find the problem and solution, which can be tedious in a large feature model. It would be easy to blame a problem on the recently added constraint that triggered the inconsistency. Still, it may be that the newly added constraint is correct and has exposed previously unnoticed bugs. Determining this would require the user to trace all the constraints involved in the problem back to the root feature. In a large feature model consisting of hundreds or thousands of features, this would be a daunting task. It would be better to point out to the user, in a meaningful way, the constraints that contribute to the inconsistency [3]. Showing exactly the constraints involved in the conflict would allow the user to more easily find and correct the bug. Tool support is critical to the adoption of SPLE. Hence, we propose a visual editor to support feature model editing and debugging better. The tool detects all errors that cause a model to be inconsistent and provides the user with hints about what the problems are, where they are, and possible fixes. It does all of this visually within the feature model editor in a non-intrusive manner, similar to how modern IDEs hint at errors

and solutions without distracting the user. Our approach affects anyone who uses feature models to develop product-line software.

II. BACKGROUND

1. Feature Model

A *feature* is a product's characteristic, quality, or functionality. Features are organized in a feature model as a tree-like structure. The relationships between parents and children, called *structural constraints*, define commonality and variability of the parent feature and ultimately the variability of the product line. The structural constraints are:

- *Mandatory*: If the parent is in a product, then the child is also in the product and vice versa.
- *Optional*: The child may or may not be in a product regardless of whether the parent is or is not in the product.
- *Alternative*: If the parent is in a product, then exactly one of its children is also in the product.
- *Or*: If the parent is in a product, then at least one of its children is also in the product.

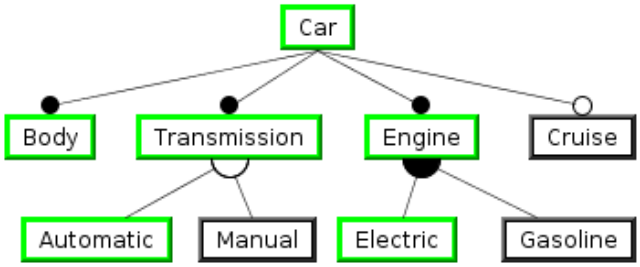


Figure 1. A feature model example of a family of cars. Features highlighted in green indicate one possible car.

When the *parent-child constraints* are insufficient, integrity constraints are used to define relationships between features that are not parent-child. They are:

- *Implies*: if *a* implies *b*, then whenever *a* is in a product, *b* must also be in the product.
- *Excludes*: if *a* and *b* exclude one another, then they can never exist together in any product.

Figure 1 is a simple feature model of a family of cars rendered in our editor. A *Mandatory* constraint is denoted by an edge with a filled small circle—for example, the edge between car and body. An *Optional* constraint is rendered as an edge with an empty circle. *Alternative* and *Or* constraints are rendered as edges with a hollow half circle and edges with a filled half circle, respectively. *Implies* constraints are rendered as unidirectional dashed edges, while *Excludes* constraints are represented as bidirectional dashed edges, as shown in Figure 2. According to the model in Figure 1, a car must have a body, an automatic or manual transmission, an engine that can be electric, gasoline, or both, and may or may not have cruise control.

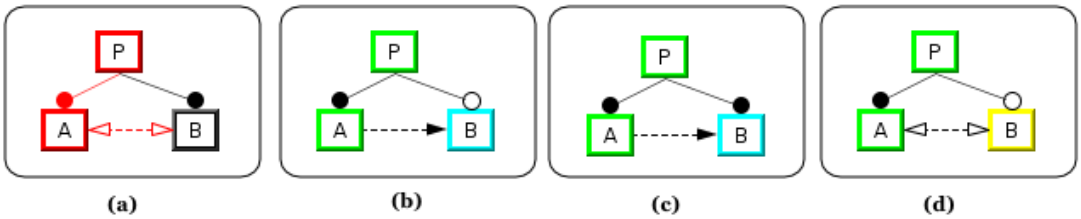


Figure 2. Problems and warnings in the feature model: (a) an inconsistent model, (b) B is a false optional feature, (c) the implication of B is redundant, (d) B is a dead feature.

Figure 2 illustrates how different problems with feature models are displayed in the editor, which currently uses color to highlight them. Figure 2(a) depicts an "inconsistent model" because *P* mandates both *A* and *B,* but *A* and *B* exclude one another; thus, no product can be derived. Figure 2(b) depicts a "*false optional feature*" warning, which is given when an *Implies* constraint effectively mandates a feature that the model labels optional. An "*implies redundancy*" warning occurs when an *Implies* constraint is not needed due to other constraints, as shown in Figure 2(c). Finally, Figure 2(d) shows a "*dead feature*", where another constraint excludes an optional feature. Although the latter three warnings do not create an inconsistent model, they may imply bad design, and thus the user should be warned.

2. Current Support for Editing and Debugging

Currently, there are fewer than a dozen feature model editors, each with a different graphical interface. Most editors provide an interface that allows users to create feature diagrams visually, with features organized into a tree-structured hierarchy. Others provide flat-tree editing, which is less useful for visualizing the feature model but still superior to textual editing. Most of these tools can detect when a model is inconsistent, and some can highlight which integrity constraints are violated [4], [5]. Most editors can detect dead features and false optional features, which are usually easy to spot and can be corrected by simply highlighting the problematic features [4]. A more difficult debugging task is correcting an inconsistent model, for which current tools provide little or no support. Current tools use propositional logic, description logic, and ad hoc methods [6], including XML Schema for validating feature models [7]. To determine if a feature model is consistent or not, it is converted to a propositional formula [4], [8], [9] or description logic [10], [11] and fed to a SAT solver or a description logic reasoner. If it is not satisfiable, then the model is inconsistent.

In general, the error reporting interfaces of current tools are weak. For example, when a new constraint is added to FeatureIDE [4] that makes the feature model inconsistent, the editor will highlight the newly added rule. However, the conflicting constraints are not highlighted, and any highlighting of previously added inconsistent constraints will be lost. Description logic systems can generate compiler-like textual error messages. Still, these messages are expressed in description logic rather than in the feature model diagram the user has been working with [11].

III. APPROACH AND CONTRIBUTION

Our approach uses propositional logic but leverages the concept of minimum unsatisfiable cores [12] and the QuickExplain algorithm [13] to identify conflicting relations and features, and then proposes possible solutions.

We adopt a similar approach to that of the Eclipse plugin dependency management system [14] for detecting errors in plugin dependencies and providing meaningful explanations to the user. This system is scalable even when faced with 10000 installable units [14]. The dependency management technology in Eclipse is also

used in Maven3 and the repository manager Nexus [14]. In addition, a feature model is a relatively easy problem for a SAT solver [8].

1. Feature Model to CNF Formula

To determine if a feature model is consistent or not, we convert a feature model to a CNF boolean formula, for which an off-the-shelf SAT solver can be used to determine the formula's satisfiability. To generate the formula, a depth-first traversal is performed. For each feature and its children, CNF clauses are generated to encode the feature's constraints and its integrity constraints. (and) Boolean operators join together all the clauses. The root feature is encoded as a single literal clause. These are the mappings of features and their constraints to CNF clauses:

- r is root: (r)
- c **mandatory** child of p: $(\neg p \vee c) \wedge (\neg c \vee p)$
- c **optional** child of p: $(\neg c \vee p)$
- $c_1, c_2, \cdots, c_n$ **or** children of p:
  $(\neg p \vee c_1 \vee c_2 \vee \cdots \vee c_n) \wedge (\neg c_1 \vee p) \wedge (\neg c_2 \vee p) \wedge \cdots \wedge (\neg c_n \vee p)$
- $c_1, c_2, \cdots, c_n$ **alternative** children of p:
  $(c_1 \vee c_2 \vee \cdots c_n \vee \neg p) \wedge (\neg c_1 \vee \neg c_2) \wedge \cdots \wedge (\neg c_1 \vee \neg c_n) \wedge (\neg c_1 \vee p) \wedge (\neg c_2 \vee \neg c_3) \wedge \cdots \wedge$
  $(\neg c_2 \vee \neg c_n) \wedge (\neg c_2 \vee p) \wedge (\neg c_{n-1} \vee \neg c_n) \wedge (\neg c_{n-1} \vee p) \wedge (\neg c_n \vee c)$
- a **implies** b: $(\neg a \vee b)$
- a **excludes** b: $(\neg a \vee \neg b)$

If a SAT solver evaluates the CNF formula to be satisfied, then the model is consistent, and inconsistent otherwise. Editors that use propositional formulas for inconsistency checking stop their debugging support here, leaving the user to identify errors and find solutions. Although our approach uses propositional formulas, we are interested in identifying the clauses responsible for the inconsistency, which is where we can provide visual debugging.

2. Supporting Debugging

Error explanation is key in helping users to understand and fix errors. To support error explanation, we need to find the minimum number of violated constraints [3] and the least number of modifications needed to correct the violation [14]. This is a problem of finding the minimum unsatisfiable cores (MUS) [12], the smallest set of clauses that makes the formula unsatisfiable, such that the removal of any clause from the set causes the formula to be satisfiable.

The MUS yields one error that contributes to a feature model's inconsistency. But it is not useful to indicate only one error when there are many. Rather, the user should be able to review all errors to plan which to fix and in what order. To detect all errors in the model, we need to remove one clause representing an integrity constraint that caused the conflict, without changing the feature model tree, and then find the next MUS to determine the next conflict. We repeat this until the formula is satisfiable, by which point all errors have been detected. This entire process is described in Algorithm 1.

1	**input** : *model*: feature model
2	**output**: *conflicts*: conflicts
3	*formula* ← toCNF(*model*);
4	**while** ¬ isSatisfiable(*formula*) do
5	

```
    mus ← getMUS(formula);
    foreach clause ∈ mus do
      conflicts ← conflicts + clause;
      if isExcludesRelation(clause) then
        formula ← formula − clause;
      end
    end
end
return conflicts;
```

Algorithm 1: Extracting conflicts in a model

Once the editor has collected all possible errors, the conflict visualization conveys as much information as possible by tracing all conflicting constraints back to the root feature and indicating which constraints and features are involved in the conflict. Any constraint in this trace is a solution; the conflict can be removed by simply removing or changing conflicting constraints. If a user adds a constraint and causes the model to be inconsistent, but the constraint is correct, then there must have been a bug in the model before adding the constraint. An example of a bug is an optional constraint created as a mandatory constraint. The constraints trace lets the user visually inspect the constraints involved in the conflict to determine which is the bug.

Figure 3 shows two errors and three warnings in the model within our editor. Errors are highlighted in red. One error is *A-B-G-F-C-A*, and another is A-H-J-K-I-H. Excluded constraints cause both, but any highlighted constraint is a possible bug and a possible solution. The warnings include *Untitled*, *D*, and *L*. *Untitled* is a dead feature because it has no parent and would never be in any product. *D* is excluded by *F,* and so *D* could never be any product. The *Implies* constraint between *G* and *L* is not needed because *L* is included in all products.

In the first error, changing any of the *Mandatory* constraints highlighted to an *Optional* constraint would remove the first error. Alternatively, the *Excludes* constraint between *G* and *F* could be removed. Removing the second error follows the same idea. Only someone with domain knowledge can decide what action to take, just as fixing a program bug requires semantic knowledge of the problem.

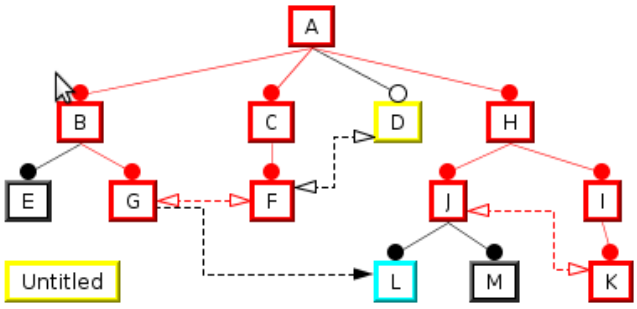


Figure 3. Editor: highlighting errors and warnings in the model.

IV. IMPLEMENTATION

The visualization of the feature model and constraints is implemented using NetBeans Visual Library [15]. The Visual Library provides basic widgets and edges for rendering an underlying model. In our case, we customized the node widgets, anchors, anchor shapes, and edge connections. The underlying models of our data structure are both a graph and a tree. A graph models the entire feature model, including constraints, while the tree data structure models the parent- child hierarchy of features. We take this approach because, in the future, we want to provide a tree rendering of the model as an alternative editing mode and for product configuration. The graph model allows greater flexibility in labeling every edge. The SAT solver we use is Sat4j [16], an open-source library that enables Java programmers to access cross-platform SAT solvers. It is used by many projects, such as FeatureIDE [4] and in Eclipse's plugin management system [14].

V. CONCLUSIONS AND FUTURE WORK

We have presented a visual feature model editor that non-intrusively hints at errors and provides a natural debugging and editing workflow. It uses the concept of minimal unsatisfiable cores to identify the constraints and features involved in a model's inconsistency and to determine the minimal change needed to fix errors visually. Other tools only report that the feature model is inconsistent. They do not provide help on how to fix the model, do not report all the errors a model may have, nor do they help the user spot bugs far from the constraints that have been causing the issue of inconsistency. Improved tool support for feature model editing and debugging will help domain and application engineers create and use feature models, thereby encouraging the adoption of feature models and SPLE as a whole.

We plan to evaluate our prototype for scalability and usability on realistic feature models; extend the tool to support cardinality-based feature model editing and debugging, product configuration, and feature model evolution; and support collaborative editing. Visual editing of complex Boolean constraints, other than implies and excludes, is in progress.

REFERENCES

[1] P. Clements and L. M. Northrop, Software Product Lines: Practices and Patterns. Addison-Wesley, 2002.
[2] D. Weiss and R. C. T. Lai, Software Product Line Engineering. Addison-Wesley, 1999.
[3] D. Batory, D. Benavides, and A. Ruiz-Cortes, "Automated analysis of feature models: challenges ahead," Commun. ACM, vol. 49, no. 12, pp. 45–47, December 2006. [Online]. Available: http://doi.acm.org/10.1145/1183236.1183264
[4] C. Kastner, T. Thum, G. Saake, J. Feigenspan, T. Leich, F. Wielgorz, and S. Apel, "FeatureIDE: A tool framework for feature-oriented software development," in Proceedings of ICSE '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 611–614. [Online]. Available: http://dx.doi.org/10.1109/ICSE.2009.5070568
[5] D. Beuche, "Modeling and building software product lines with pure::variants," in Proceedings of the 15th International Software Product Line Conference, Volume 2, ser. SPLC '11. New York, NY, USA: ACM, 2011, pp. 46–1. [Online]. Available: http://doi.acm.org/10.1145/2019136.2019190
[6] D. Benavides, S. Segura, and A. Ruiz-Cortés, "Automated analysis of feature models 20 years later: A literature review," Inf. Syst., vol. 35, no. 6, pp. 615–636, September 2010. [Online]. Available: http://dx.doi.org/10.1016/j.is.2010.01.001
[7] V. Cechticky, A. Pasetti, O. Rohlik, and W. Schaufelberger, "XML-based feature modeling," in Software Reuse: Methods, Techniques and Tools: 8th International Conference, ICSR 2004. Springer-Verlag, 2004, pp. 5–9.
[8] M. Mendonca, M. Branco, and D. Cowan, "S.P.L.O.T.: software product lines online tools," in Proceedings of the 24th ACM SIGPLAN conference companion on Object-oriented programming

systems, languages and applications, ser. OOPSLA '09. New York, NY, USA: ACM, 2009, pp. 761–762. [Online]. Available: http://doi.acm.org/10.1145/1639950.1640002

[9] M. Antkiewicz and K. Czarnecki, "FeaturePlugin: feature modeling plug-in for Eclipse," in Proceedings of the OOPSLA workshop on Eclipse Technology Exchange, ser. eclipse '04. New York, NY, USA: ACM, 2004, pp. 67–72. [Online]. Available: http://doi.acm.org/10.1145/1066129.1066143

[10] M. Noorian, A. Ensan, E. Bagheri, H. Boley, and Y. Biletskiy, "Feature model debugging based on description logic reasoning," in The 17th International Conference on Distributed Multimedia Systems (DMS 2011). KSI, 2011.

[11] H. H. Wang, Y. F. Li, J. Sun, H. Zhang, and J. Pan, "Verifying feature models using OWL," Web Semant., vol. 5, no. 2, pp. 117–129, June 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1265608.1265750

[12] I. Lynce and J. P. Marques-Silva, "On computing minimum unsatisfiable cores," in International Symposium on Theory and Applications of Satisfiability Testing, May 2004, pp. 305– 310. [Online]. Available: http://eprints.ecs.soton.ac.uk/12252/

[13] U. Junker, "QUICKXPLAIN: preferred explanations and relaxations for over-constrained problems," in Proceedings of AAAI '04. AAAI Press, 2004, pp. 167–172. [Online]. Available: http://dl.acm.org/citation.cfm?id=1597148.1597177

[14] D. Le Berre and P. Rapicault, "Dependency management for the Eclipse ecosystem: Eclipse p2, metadata and resolution," in Proceedings of the 1st international workshop on Open component ecosystems, ser. IWOCE '09. New York, NY, USA: ACM, 2009, pp. 21–30. [Online]. Available: http://doi.acm.org/10.1145/1595800.1595805

[15] "Netbeans visual library 2.0," 2011. [Online]. Available: http://platform.netbeans.org/graph/

[16] D. Le Berre and A. Parrain, "The Sat4j li- brary 2.2, system description," System, vol. 7, no. June 2009, pp. 59–64, 2010. [Online]. Available: http://jsat.ewi.tudelft.nl/content/sd/JSAT7 4 LeBerre.pdf

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Issue Analyzer: Similarity-Based Grouping and Team Routing for Mobile Crash Logs

Selim Toksal
Technology Department
Trendyol Group
Istanbul, Türkiye
selim.toksal@trendyol.com

Berkay Mutlu
Technology Department
Trendyol Group
Istanbul, Türkiye
berkay.mutlu@trendyol.com

Recep Çelik
Technology Department
Trendyol Group
Istanbul, Türkiye
recep.celik@trendyol.com

Tuna Çakar
Computer Engineering
MEF University
Istanbul, Türkiye
cakart@mef.edu.tr

Abstract
This study presents an Issue Analyzer for centralized analysis of mobile errors and crash logs in large software organizations. The system maps heterogeneous records to a canonical schema, performs build-aware symbolication and normalization, represents each issue through message, stack-trace, and contextual views, and groups related records by a mathematically defined composite similarity score. The reference implementation uses a Sentence-BERT message encoder, cosine similarity, weighted stack-frame overlap, normalized longest-common-subsequence similarity, and a gradient-boosted routing model. Candidate teams are returned with confidence scores, supporting evidence, and an option to abstain. To reduce historical routing bias, training labels are based on the final resolving team rather than the initial assignee, while reassignment histories and human corrections are retained as audit signals. Offline routing results are complemented by a controlled-replay benchmark and a limited, non-randomized shadow-mode workflow evaluation; neither is treated as definitive production evidence. The approach positions issue analysis as an explainable, human-in-the-loop decision-support capability accessible through a Slack-like conversational interface.

Keywords: mobile application, crash logs, log normalization, crash clustering, bug triage, team routing, explainability, conversational analytics.

I.INTRODUCTION

Mobile applications are primary service channels for e-commerce, payment, logistics, communication, and content platforms. A recurring crash can interrupt a transaction, increase support effort, and weaken confidence in releases. In large engineering organizations, however, crash and error records are distributed across monitoring tools, applications, versions, devices, ticket systems, and internal data stores. Similar failures are reported repeatedly with different variable values or partially different traces, and the first team assignment often depends on individual experience. The proposed Issue Analyzer converts these fragmented records into normalized issue families and evidence-backed routing recommendations. Unlike classification frameworks that treat a predicted label as the final output, the approach is designed around operational use: users can inspect similar historical cases, understand why a team was suggested, correct the result, and query release-level patterns through a Slack-like interface. This design extends access beyond developers to stakeholders in quality, product, and operations. The main contributions are fourfold. First, the study defines a build-aware normalization process that uses Android R8/ProGuard mapping files and iOS symbol files before frame extraction. Second, it specifies

the exact representations, distance measures, and mathematical formulation used for multi-view issue similarity. Third, it introduces routing safeguards that learn from the final resolving team, audit reassignment paths, and abstain when evidence is weak or conflicting. Fourth, it separates offline model evaluation from production workflow and scalability validation, avoiding unsupported production claims.

II.RELATED WORK

Bug-triage research has shown that historical assignments can support semi-automated fixer recommendation [1], while bug-tossing graphs and data-reduction methods can improve routing quality and reduce noisy training evidence [2], [3]. Studies on report quality, duplicate detection, and mobile bug-report completion further show that text alone is insufficient when technical context is missing or inconsistent [4]–[7]. Crash bucketing and localization methods therefore exploit stack structure and execution context [8]–[12]. The log-analysis literature provides complementary methods for template extraction, clustering, anomaly detection, and evaluation [13]–[20]. Sentence-BERT offers compact semantic representations for short issue messages [21], and gradient-boosted trees provide a practical supervised model for mixed numerical and categorical routing features [22]. The present work combines these ideas in a workflow-centered mobile issue-analysis architecture with explicit human feedback and conversational access.

III.SYSTEM ARCHITECTURE AND METHOD

1. Architecture and Canonical Issue Schema

Historical backfill and streaming events are validated and mapped to a canonical record that contains the raw and normalized message, exception class, original and symbolicated frames, application and build identifiers, platform, device family, operating system version, timestamp, release window, feature/module hints, source link, and data quality flags. The same schema is used for training, cluster construction, online scoring, and audit. Figure 1 shows the end-to-end architecture.

2. Build-Aware Normalization and Ambiguous Traces

Normalization first replaces unstable values with typed placeholders such as <UUID>, <HEX>, <INT>, <URL>, <PATH>, <TIME>, and <TOKEN>. The system retains both the raw text and the normalized template so that information is not irreversibly lost. Token replacement is applied only after source-specific parsing; for example, line numbers and addresses are masked, while exception classes, method names, package paths, and stable error codes are preserved.

Table 1. Core Pipeline Layers And Outputs

Layer	Main operation	Output
Ingestion	Validate backfill and streaming mobile issue events	Canonical raw events
Normalization	Symbolicate, parse, mask unstable tokens, and rank frames	Stable issue representation
Grouping	Compute multi-view similarity and update clusters	Issue families and profiles

Layer	Main operation	Output
Routing	Rank candidate teams, score confidence, and abstain	Explainable team suggestions
Access/feedback	Answer grounded queries and record corrections	Operational insight and audit signals

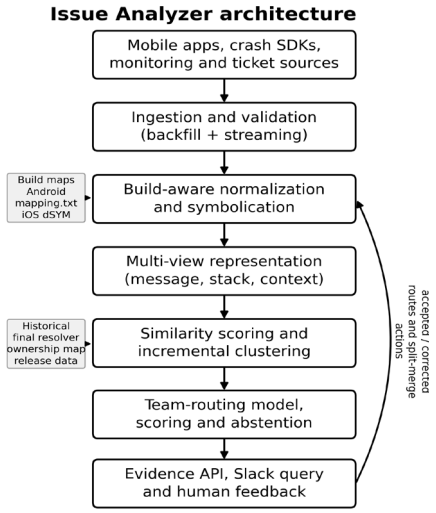


Figure 1. Build-aware Issue Analyzer architecture and human-feedback loop.

Frame extraction is build-aware. For Android, the application identifier, version code, and build identifier select the corresponding R8/ProGuard mapping.txt file; retracing is completed before package parsing and library filtering. For iOS, the bundle identifier, build number, architecture, and binary UUID select the matching dSYM and, where applicable, BCSymbolMap. Each frame stores the raw symbol, mapped symbol, mapping-file version, application-versus-library flag, and symbolication status. Missing or conflicting mappings are not silently accepted: the record is marked as partially symbolicated and its stack contribution is down-weighted.

Mobile traces can remain ambiguous due to obfuscation, asynchronous callbacks, shared framework frames, truncation, or different execution paths that lead to the same exception. The analyzer, therefore, does not use a single trace hash as a universal identifier. It separates records by application/build/ABI when necessary, uses only the top K application frames for ordered comparison, and avoids deterministic merging when symbolication confidence is low. In such cases, message semantics, exception class, release context, and module evidence receive greater weight; low-margin cases are sent for manual split-or-merge review.

3. Similarity Representation and Grouping

The reference configuration uses the 384-dimensional all-MiniLM-L6-v2 Sentence-BERT encoder for normalized messages [21]. Embeddings are L2-normalized and compared by cosine similarity. A character 3–

5-gram TF–IDF vector is retained as a lexical safeguard for short error templates and product-specific codes. The message score is

$$s_msg(i,j) = \rho\, cos(e_i,e_j) + (1-\rho)\, cos(v_i,v_j), \quad 0 \le \rho \le 1 \quad (1)$$

where e is the Sentence-BERT vector and v is the TF–IDF vector. Stack similarity combines weighted frame overlap and order. Let F_i be the ordered top-K application frames, and let ak = 1/log2 (k+1) be the rank weight. The stack score is

$$s_stack(i,j) = \alpha\, J_w(F_i,F_j) + (1-\alpha)\, LCS(F_i,F_j)/max(|F_i|,|F_j|) \quad (2)$$

J_w is weighted Jaccard similarity, and LCS is the longest common subsequence. Context similarity compares application, platform, build, version, environment, device family, and release window. Exact categorical matches are combined with an exponential time-decay kernel:

$$s_ctx(i,j) = [\Sigma_k\, q_k\, I(x_{ik}=x_{jk}) + q_t\, exp(-|t_i-t_j|/h)] / [\Sigma_k\, q_k + q_t] \quad (3)$$

The final issue-pair score is a convex combination of message, stack, context, and module evidence:

$$S(i,j) = w_m\, s_{msg} + w_s\, s_{stack} + w_c\, s_{ctx} + w_o\, s_{owner}, \quad \Sigma_w=1,\ w\ge 0 \quad (4)$$

The weights and merge threshold are selected only on validation data by maximizing pairwise F1 subject to a maximum acceptable over-merge rate. The selected validation values were w_m=0.30, w_s=0.40, w_c=0.15, w_o=0.15, ρ=0.75, α=0.60, and τ_{merge}=0.82. A record joins the highest-scoring compatible cluster only when $S \ge \tau_{merge}$ and no hard conflict exists in application, build, or exception family; otherwise, a new cluster is opened. Manual split and merge decisions are versioned as feedback.

4. Team Routing, Explainability, and Bias Control

Routing is formulated as a ranked team recommendation rather than a forced assignment. The supervised component is an XGBoost multiclass classifier [22] that uses cluster-level features: top-message and stack similarities to resolved clusters, exception and module indicators, application/build/platform metadata, release-change overlap, frequency trend, and ownership-map compatibility. Model confidence is reported without post-hoc calibration; the abstention thresholds are applied directly to the model output. A final score combines the model probability with independent historical, ownership-map, and release evidence:

$$R(t|c) = \eta_p\, p_{XGB}(t|c) + \eta_h\, h(t|c) + \eta_m\, m(t|c) + \eta_r\, r(t|c) - \lambda\, b(t|c) \quad (5)$$

Here, h is similarity-weighted evidence from previously resolved clusters, m is module-map compatibility, r is release-change evidence, and b is a misrouting penalty derived from reassignment history. The interface returns the top-k teams, confidence, supporting traces/tickets, and the contribution of each evidence group. The system abstains when the top score is below δ, or the margin between the first two teams is below μ; the abstention thresholds were δ = 0.62 and μ = 0.10.

Table 2. Reference Representations And Similarity Functions

Signal	Representation	Comparison
Message	all-MiniLM-L6-v2 (384-D) + char 3–5-gram TF–IDF	Cosine mixture, Eq. (1)
Stack	Exception + ordered top-K symbolicated application frames	Weighted Jaccard + normalized LCS, Eq. (2)
Context	App, platform, build, version, device, release, and time	Weighted match + time kernel, Eq. (3)
Ownership	Versioned module-to-team map and resolved-ticket evidence	Probabilistic/binary compatibility

The abstention thresholds were selected empirically rather than by search over a reserved partition. During development, routing suggestions were reviewed by domain experts and marked correct or incorrect; the resulting outcomes were used to identify a confidence level at which the proportion of incorrect routed suggestions became acceptable to the operating teams, and a top-1/top-2 margin below which reviewers frequently disagreed with the ranking. The values $\delta = 0.62$ and $\mu = 0.10$ reflect the chosen operating point. Because the same judgments informed both the choice of thresholds and the observed behavior at those thresholds, the operating point is fitted to the observed data, and its selective risk on unseen data may be higher than observed. Five design measures are intended to limit historical assignment bias. Training labels use the final resolving team, not the first assignee. Tickets that were repeatedly bounced, reopened, or closed without a verified owner are excluded or given lower weight. Reassignment paths are monitored as a separate error signal rather than treated as ground truth. Ownership maps are versioned and applied based on the issue date to prevent present-day mappings from leaking into older records. Finally, per-team precision, correction rate, and confidence behavior are audited; any conflict between historical evidence and the current ownership map triggers abstention or human review rather than automatic routing. The Slack-like layer interprets a question, converts it into a restricted query over canonical issue tables and cluster profiles, and summarizes the returned evidence. The language model does not independently create ownership or diagnostic facts. Each response includes the query scope, time window, affected versions, cluster identifiers, confidence, and links to source records. Role-based views allow developers to inspect raw traces while product and operations users receive concise impact summaries. Accepted and corrected routes, as well as manual cluster split/merge actions, are written back to the feedback store.

IV.EVALUATION METHODOLOGY

Table 3 reports measurements collected during the iterative development of the routing component on operational data rather than results from a held-out benchmark evaluation. The working corpus contained 48,736 issue records grouped into 2,614 issue families, collected from January to December 2025, across 10 candidate software teams. The system operated continuously over this period, and issue records were added to the training data as they were resolved, so each configuration was trained on all records resolved up to the point of its introduction. Each configuration was assessed by domain experts from the engineering organization, who reviewed the routing suggestions produced by that configuration and marked each as correct or incorrect. Precision, recall, accuracy, and F1 in Table 3 are computed from these expert judgments. This procedure reflects

how the system was actually developed and evaluated in its operating environment; its limitations are stated in Section VI. Experiments were run on a 32-core AMD EPYC 7543P CPU, an NVIDIA A10 GPU with 24 GB of memory, and 128 GB of RAM, using Ubuntu 22.04, Python 3.11, sentence-transformers 2.6.1, XGBoost 2.0.3, and FAISS 1.8.0. Grouping behavior is qualitatively characterized in this study. Pairwise precision, recall, F1, cluster purity, and over- and under-merge rates require an independently labeled reference partition and are identified as required instrumentation in Section VI. Routing is assessed using precision, recall, accuracy, and F1 computed from developer judgments, along with the observed correction rate. Expected calibration error, top-3 accuracy, per-team breakdowns, abstention coverage, and selective risk require a reserved evaluation partition with independently assigned labels and are not reported here; they are identified as required instrumentation in Section VI. The development trace in Table 3 shows the order in which message, stack, and ownership evidence were introduced. Still, it is not a controlled ablation: the training corpus grew as issues were resolved, so later configurations were trained on more data and on additional evidence sources. The contributions of the individual signals cannot be separated from the effect of the additional training data. A controlled ablation, in which each signal is removed from the full configuration while the training corpus is held fixed, has not been performed and is identified as required work in Section VI. Workflow validation requires time-to-first-owner, number of reassignments, analyst effort, and user-rated explanation usefulness; these outcomes cannot be inferred from offline classification scores.

1. Scalability and Workflow Validation

A controlled replay benchmark was conducted with events that preserved the observed distributions of message length, stack depth, build, platform, and cluster size. Each test used a five-minute warm-up and three measured repetitions; Table 4 reports the median p50, p95, and p99 values across repetitions. A two-week shadow-mode evaluation was also used to compare routing workflow indicators without automatically changing ticket ownership. The shadow-mode comparison was observational rather than randomized and therefore supports only descriptive comparison. These results characterize the stated test environment and should not be interpreted as a guaranteed production service-level agreement. Each iteration adds one source of evidence to the preceding configuration, and the routing model is XGBoost throughout. Four properties of this assessment should be noted. First, the training corpus grew continuously as issues were resolved, so later configurations were trained on more data than earlier ones, and the effect of the added evidence cannot be separated from that of the added data. Second, the configurations were reviewed sequentially during development rather than simultaneously on a fixed record set, so the issues reviewed differ across rows, and the values are not paired measurements. Third, reviewers were not blind to the routing suggestion when making their judgment. Fourth, no held-out partition was reserved. We therefore read Table 3 as a record of the system's development trajectory rather than as a controlled comparison between configurations.

Table 3. High-Load And Workflow Measurements

Iteration	**Message representation**	**Stack evidence**	**Ownership evidence**	**Precision**	**Recall**	**Acc.**	**F1**
1	Char 3–5-gram TF–IDF over crash title and error message	-	-	0.55	0.48	0.52	0.51

Iteration	Message representation	Stack evidence	Ownership evidence	Precision	Recall	Acc.	F1
2	+ Sentence-BERT over the same fields	-	-	0.67	0.71	0.69	0.69
3	as Iteration 2	Weighted frame overlap + normalized LCS	-	0.72	0.68	0.70	0.70
4	as Iteration 3	as Iteration 3	Versioned ownership-map compatibility	0.86	0.89	0.87	0.87

Across the four configurations in Table 3, expert-reviewed routing precision rose from 0.55 to 0.86, recall from 0.48 to 0.89, and F1 from 0.51 to 0.87. The largest single increments followed the introduction of the Sentence-BERT message encoder and the ownership-map evidence. Because the configurations were reviewed sequentially on different issues, the training corpus grew over the same period, and no held-out partition was used, these values reflect the system's development trajectory rather than its expected accuracy on unseen data. They cannot be compared directly with published benchmark figures. In the controlled replay, the system sustained 165 events/s, with p95 ingestion and incremental-cluster latencies of 34 ms and 57 ms, respectively; the p95 natural-language query response time was 910 ms. During the two-week shadow-mode period, median time-to-first-owner was 17 minutes, compared with 41 minutes in the preceding baseline window, and the mean number of reassignments per issue was 0.6, compared with 1.6; users corrected 12.8% of routing suggestions. Because the shadow-mode evaluation was observational rather than randomized and ran concurrently with ordinary release activity, these differences are reported as observed associations rather than as causal effects of the Issue Analyzer. These measurements provide initial evidence of workflow and scalability, but they do not establish long-term production performance across every release, traffic pattern, or organizational structure.

Table 4. High-Load And Workflow Measurements

Measurement	Reporting unit	Result
Daily ingested volume and sustained throughput	events/day; events/s	96,000/day; 165/s sustained
Ingestion-to-canonical-record latency	p50/p95/p99 ms	11/34/68 ms
Incremental cluster update latency	p50/p95/p99 ms/event	18/57/109 ms
Full cluster compilation time	s per 10k/100k events	38/412 s
Natural-language query response time	p50/p95/p99 ms	390/910/1,380 ms
Embedding and index footprint	CPU %, RAM GB, GPU GB, index GB	36%; 10.8 GB; 4.7 GB; 0.76 GB

Measurement	Reporting unit	Result
Workflow indicators (shadow period vs. baseline window, non-randomized)	time-to-owner; reassignments; correction rate	41 vs. 17 min; 1.6 vs. 0.6; 12.8%

Qualitatively, the architecture is designed to address a common usability gap. Stakeholders can work with issue-family summaries instead of isolated logs, inspect release-level growth, and receive ranked teams with evidence rather than an unexplained label. Build-aware symbolication and abstention are particularly important in mobile settings because variable replacement alone cannot resolve ambiguous or partially mapped traces.

V.DISCUSSION

The Issue Analyzer extends classical bug triage by combining technical similarity with the operational path from evidence to action. The classifier is not treated as an end state: users see comparable clusters, current ownership metadata, confidence scores, alternative teams, and the reason for abstention. Corrections become versioned signals for later model and rule updates. We argue that this human-in-the-loop structure is preferable to a black-box auto-assignment mechanism, especially during incidents or organizational changes; this is a design position rather than an empirically tested comparison. The proposed safeguards are intended to reduce, but cannot eliminate, historical bias; their actual effect has not been quantified in this study. Even the final resolver may reflect organizational constraints rather than ideal ownership, and teams with more tickets may dominate the training data. Regular label audits, class-balanced learning, per-team confidence auditing, recency weighting, and current ownership-map checks are therefore required. Similarly, multi-view similarity cannot guarantee root-cause equivalence. Low-confidence symbolication, shared library frames, and asynchronous execution can still produce over-merging or under-merging; split/merge review and uncertainty reporting remain necessary. The main limitation is external validity. The corpus represents one organization and one annual period, the load results come from a controlled replay rather than unrestricted production traffic, and the workflow comparison was a two-week shadow-mode evaluation rather than a randomized field experiment. Future work should repeat the evaluation across applications and organizations, compare routing against initial-assignee and current-ownership baselines, report ablations and confidence intervals, and conduct a longer controlled pilot using time-to-owner, reassignment, and user-effort outcomes. The measurements in this study were collected as part of the system's ongoing operation rather than as a reproducible experimental artifact. Engineers reviewed routing suggestions during normal work, and the resulting correct/incorrect judgments were recorded as operational indicators; the intermediate evaluation records from that period were not retained for later re-analysis. The system itself remains operational, but the specific record sets underlying Tables 3 and 4 cannot be reconstructed. Several evaluations described in Section IV are therefore not reported here. Specifically, we do not report a controlled ablation with a fixed training corpus, a comparison against non-learned routing baselines such as historical-majority or ownership-map assignment, pairwise grouping metrics against an independently labeled reference, per-team performance breakdowns, expected calibration error, abstention coverage, and selective risk, confidence intervals, or a sensitivity analysis over δ and μ. Each of these requires measurement on a reserved partition with independently assigned labels, which the available records do not support. The consequence is that the quantitative values in this paper should be read as operational observations from a single organization over a single 12-month period, not as benchmark results. They indicate the trajectory the system followed during development and the workflow indicators observed during the shadow period. Still, they do not establish how the approach would perform on unseen data or how it compares with alternative designs. Because the system remains operational, a prospective evaluation on a chronologically reserved partition, with labels assigned independently of the system's output, is feasible and is the primary direction for future work.

VI.CONCLUSION

This study presented a centralized Issue Analyzer for mobile crash and error records. The revised method specifies build-aware symbolication, typed normalization, exact multi-view similarity equations, a Sentence-BERT message representation, stack-aware comparison, XGBoost-based team ranking, abstention, historical-bias controls, grounded Slack-style access, and a human-feedback loop. These elements are intended to make the approach operationally useful without turning routing into an unsupported automatic decision. The available development measurements and the controlled-load benchmark indicate that the architecture is technically feasible in the evaluated setting, and the shadow-mode indicators are favorable. Because that comparison was not randomized and the underlying evaluation records were not retained for re-analysis, these observations do not establish that the Issue Analyzer caused the observed workflow differences. They do not, however, constitute a universal guarantee of production scalability. Broader longitudinal and multi-organization validation is required before the reported thresholds and capacity values can be generalized. Within these limits, the Issue Analyzer provides a reproducible quality-intelligence layer for centralized mobile issue analysis and evidence-based team routing.

REFERENCES

[1] J. Anvik, L. Hiew, and G. C. Murphy, “Who should fix this bug?” in Proc. 28th Int. Conf. Software Engineering, 2006, pp. 361–370.

[2] G. Jeong, S. Kim, and T. Zimmermann, “Improving bug triage with bug tossing graphs,” in Proc. 7th Joint Meeting of the European Software Engineering Conference and ACM SIGSOFT Symposium Foundations of Software Engineering, 2009, pp. 111–120.

[3] J. Xuan, H. Jiang, Y. Hu, Z. Ren, W. Zou, Z. Luo, and X. Wu, “Towards effective bug triage with software data reduction techniques,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 1, pp. 264–280, 2015.

[4] T. Zimmermann, R. Premraj, N. Bettenburg, S. Just, A. Schroeter, and C. Weiss, “What makes a good bug report?” IEEE Trans. Software Eng., vol. 36, no. 5, pp. 618–643, 2010.

[5] P. Runeson, M. Alexandersson, and O. Nyholm, “Detection of duplicate defect reports using natural language processing,” in Proc. 29th Int. Conf. Software Engineering, 2007, pp. 499–508.

[6] K. Moran, M. Linares-Vásquez, C. Bernal-Cárdenas, and D. Poshyvanyk, “Auto-completing bug reports for Android applications,” in Proc. 10th Joint Meeting Foundations of Software Engineering, 2015, pp. 673–686.

[7] H. Liu, M. Shen, J. Jin, and Y. Jiang, “Automated classification of actions in bug reports of mobile apps,” in Proc. 29th ACM SIGSOFT Int. Symp. Software Testing and Analysis, 2020, pp. 128–140.

[8] Y. Dang, R. Wu, H. Zhang, D. Zhang, and P. Nobel, “ReBucket: A method for clustering duplicate crash reports based on call stack similarity,” in Proc. 34th Int. Conf. Software Engineering, 2012, pp. 1084–1093.

[9] R. Wu, H. Zhang, S.-C. Cheung, and S. Kim, “CrashLocator: Locating crashing faults based on crash stacks,” in Proc. Int. Symp. Software Testing and Analysis, 2014, pp. 204–214.

[10] M. Gomez, R. Rouvoy, B. Adams, and L. Seinturier, “Reproducing context-sensitive crashes of mobile apps using crowdsourced monitoring,” in Proc. Int. Conf. Mobile Software Engineering and Systems, 2016, pp. 88–99.

[11] S. Feng and C. Chen, “Prompting is all you need: Automated Android bug replay with large language models,” in Proc. 46th IEEE/ACM Int. Conf. Software Engineering, 2024. pp. 803–815

[12] Y. Huang et al., “CrashTranslator: Automatically reproducing mobile application crashes directly from stack trace,” in Proc. 46th IEEE/ACM Int. Conf. Software Engineering, 2024, pp. 190–202.

[13] M. Du, F. Li, G. Zheng, and V. Srikumar, “DeepLog: Anomaly detection and diagnosis from system logs through deep learning,” in Proc. ACM SIGSAC Conf. Computer and Communications Security, 2017, pp. 1285–1298.

[14] R. Vaarandi and M. Pihelgas, “LogCluster: A data clustering and pattern mining algorithm for event logs,” in Proc. 11th Int. Conf. Network and Service Management, 2015, pp. 1–7.

[15] P. He, J. Zhu, Z. Zheng, and M. R. Lyu, “Drain: An online log parsing approach with fixed depth tree,” in Proc. IEEE Int. Conf. Web Services, 2017, pp. 33–40.

[16] S. He, J. Zhu, P. He, and M. R. Lyu, “Experience report: System log analysis for anomaly detection,” in Proc. IEEE 27th Int. Symp. Software Reliability Engineering, 2016, pp. 207–218.

[17] J. Zhu, S. He, P. He, J. Liu, and M. R. Lyu, "Loghub: A large collection of system log datasets for AI-driven log analytics," in Proc. IEEE 34th Int. Symp. Software Reliability Engineering (ISSRE), 2023, pp. 355–366.

[18] S. Hashemi and M. Mäntylä, “OneLog: Towards end-to-end software log anomaly detection,” Automated Software Engineering, vol. 31, no. 2, Art. no. 37, 2024.

[19] T. Cui et al., "LogEval: A comprehensive benchmark suite for LLMs in log analysis," Empirical Software Engineering, vol. 30, no. 6, Art. no. 173, 2025.

[20] Z. A. Khan, D. Shin, D. Bianculli, and L. C. Briand, "Impact of log parsing on deep learning-based anomaly detection," Empirical Software Engineering, vol. 29, no. 6, Art. no. 139, 2024.

[21] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing, 2019, pp. 3982–3992.

[22] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Design Science Research for Digital Platforms in Resource-Constrained Environments: The Lumikiza Case

Morley Mujansi
Department of Information Systems
Copperbelt University
Kitwe, Zambia
morley020@gmail.com
ORCID: 0000-0001-6513-6388

Jeremiah J Mwiinga
Department of Health Sciences
Namibia University of Science and Technology
Windhoek, Namibia
jjmwiinga@gmail.com
ORCID: 0009-0007-5483-4966

Alice P.S Shemi
Department of Information Systems
Copperbelt University Kitwe, Zambia
sheiap@gmail.com
ORCID: 0000-0003-2951-7743

Lilian Mungulube
School of Business Studies, Zambian Open University
Lusaka, Zambia
lillianmungulube@gmail.com
ORCID: 0009-0002-6594-0725

Abstract
The application of Design Science Research (DSR) in resource-constrained environments presents unique challenges distinct from those in well-resourced research settings. This paper presents a methodological framework for conducting DSR in developing-economy contexts, illustrated through the development and evaluation of Lumikiza, a digital advisory-matching platform for Zambian Small and Medium Enterprises (SMEs). Using a sequential exploratory mixed-methods design, we show how the three DSR cycles (relevance, design, and rigor) were operationalized in a context marked by limited digital infrastructure, trust deficits, and regulatory complexity. Ten purposively sampled advisory experts (semi-structured interviews) informed the design of a transparent Weighted Multi-Criteria Decision Analysis (MCDA) matching engine, which was then evaluated with 200 SME founders using an integrated Task-Technology Fit (TTF), UTAUT, and DeLone and McLean instrument (Cronbach's α = 0.69–0.82 across constructs). The overall evaluation model was not statistically significant (R^2 = 0.066, F = 1.95, p = 0.064); within this weak model, Information Quality was the only significant predictor of Net Benefits (β = 0.167, p = 0.038, f^2 = 0.071). We interpret this as a preliminary, context-specific signal which is a tentative "Utilitarian Priority Pattern" rather than an established finding. Methodological contributions include: (1) a framework for adapting expert interviews to inform MCDA-based design in low-data environments, (2) strategies for maintaining theoretical rigor in scenario-based prototype evaluation, and (3) approaches for integrating qualitative and quantitative data across DSR cycles, with implications for SDG 8 (Decent Work and Economic Growth) and SDG 9 (Industry, Innovation and Infrastructure).

Keywords: Design Science Research, Methodology, Developing Economies, Digital Platforms, SMEs, Applied Computing

I.INTRODUCTION

Design Science Research (DSR) has emerged as a dominant methodological paradigm within Information Systems research for creating and evaluating purposeful artifacts that address real-world problems [1]. Unlike behavioral research, which seeks to explain existing phenomena, DSR is oriented toward generating knowledge

through the systematic creation and evaluation of artifacts. This process yields both practical solutions and theoretical contributions [2].

While DSR has been widely applied in well-resourced research settings, its application in resource-constrained environments—particularly in developing economies—presents distinctive methodological challenges that remain underexplored in the literature. These challenges include limited access to participants, unreliable digital infrastructure, trust deficits in digital platforms, regulatory complexity, and the need to balance theoretical rigor with practical utility in contexts with scarce research resources. This paper addresses the following research question: How can the DSR methodology be effectively operationalized for digital platform development in resource-constrained developing-economy contexts? We position this contribution carefully. Individually, purposive expert sampling, open-source development, scenario-based evaluation, and effect-size reporting are established research practices rather than novel inventions in themselves [3]. What this paper contributes is not a new method in isolation, but (a) a specific combination and sequencing of these practices into an operationalized framework for DSR in resource-constrained developing-economy contexts, and (b) an instrumental case, Lumikiza, through which that combination is demonstrated and critically examined, including where the evaluation did not perform as strongly as hoped. We treat the framework's elements as propositions for testing in other resource-constrained contexts rather than as settled, generalized guidance. Drawing on the development of this digital advisory matching platform for Small and Medium Enterprises (SMEs) in Zambia, we present a methodological framework that addresses the distinctive challenges of conducting DSR in resource-constrained settings. The paper contributes:

- A framework for adapting expert interviews to inform Multi-Criteria Decision Analysis (MCDA) weighting in low-data environments
- Strategies for maintaining theoretical rigor in scenario-based prototype evaluation
- Approaches for integrating qualitative and quantitative data in DSR cycles
- Practical guidance for researchers developing digital platforms in similar contexts

The paper is structured as follows: a review of relevant literature; a methodological framework; an application of the framework through the Lumikiza case study, including its sampling, instrumentation, and ethical procedures; results; a discussion of findings; limitations; contributions; and conclusion.

II.LITERATURE REVIEW

1. Design Science Research in Information Systems

Design Science Research occupies a distinct methodological position within Information Systems. Where behavioral research seeks to explain phenomena that already exist, DSR is oriented toward creating something new and evaluating its effectiveness [1]. The methodology is structured around three overlapping cycles: relevance, design, and rigor. The relevance cycle grounds the research in real-world practice, ensuring that the artifact addresses an actual organizational problem and is evaluated in an appropriate context. The design cycle guides iterative development, involving the construction and evaluation of artifacts. The rigor cycle ensures theoretical accountability, anchoring evaluation in established theoretical constructs rather than ad hoc criteria [1]. Reference [2] distinguishes four types of DSR contributions based on artifact novelty and knowledge contribution: invention, improvement, exaptation, and routine design. Recent scholarship further cautions that DSR contributions are frequently judged narrowly on the artifact produced, when equal attention is warranted to the theoretical knowledge generated [3]—a balance we return to in Section VI when weighing what this case does and does not establish. A defining characteristic of DSR is the status it accords to the artifact itself. Rather

than treating the prototype as a means to an end, DSR positions the artifact as the primary instrument through which assumptions are tested and new knowledge is generated. Insight emerges not solely from what participants report, but from how they actually engage with a functioning system.

2. DSR in Resource-Constrained and Developing Economy Contexts

Questions have been raised in the literature regarding whether DSR can yield credible results outside of well-resourced research or commercial environments. However, recent scholarship provides substantive evidence to the contrary. Reference [4] demonstrates the successful application of DSR in the development of Ethiopian government digital platforms, while [5] reports comparable results in resource-constrained Tanzanian contexts. Reference [6] further establishes that for early-stage artifacts, a mixed-methods approach incorporating scenario-based evaluation constitutes a methodologically rigorous basis for assessment. Reference [7] similarly confirms that DSR is well-suited to the complexities of digital innovation in organizationally demanding environments. Earlier work on ICT service development in low-resource settings emphasizes iterative, context-sensitive design over the transplantation of best practices [8]. More recently, Guracho et al. [9] applied a DSR methodology to develop a mobile mental health application for a low-resource Ethiopian setting, illustrating that structured DSR activities—from problem identification through evaluation—remain applicable even when infrastructure, workforce, and data constraints depart substantially from those assumed in mainstream IS research. Despite these advances, the literature reveals persistent gaps in understanding how DSR can be effectively operationalized in resource-constrained contexts. Studies applying DSR in developing economies have focused primarily on government or health systems rather than SME services, and few have detailed the practical methodological adaptations required for low-resource settings.

3. Digital Platforms for African SMEs

Beyond DSR methodology specifically, a growing body of work examines digital platforms as vehicles for value co-creation in African entrepreneurship ecosystems. Oredo et al. [10] document the design and implementation of a digital platform supporting entrepreneurship ecosystems in sub-Saharan Africa, while Hlongwane and Grobbelaar [11] propose an ecosystem-level framework for value creation through health information platforms in South Africa. Noutchie [12] finds that digital-platform maturity and entrepreneurial digital literacy vary considerably across African economies, cautioning against assuming that design lessons transfer uniformly across the continent. Trust is a recurring theme in this literature. Mou, Shin, and Cohen [13] show that perceived trust and usefulness jointly shape consumer acceptance of e-services, consistent with the trust-centered design emphasis adopted in this study. More broadly, the digital transformation of advisory and consulting services themselves remains comparatively understudied outside high-income markets [14].

4. The Zambian SME Context

Small and Medium Enterprises comprise the vast majority of registered businesses in Zambia and are widely cited as key drivers of inclusive economic growth [15]. Despite this, SME founders frequently lack access to strategic, governance, or compliance guidance that could meaningfully improve their prospects for survival and growth [16]. The advisory gap in Zambia is characterized by: (1) reliance on informal peer networks rather than structured professional relationships [17], a pattern consistent with broader accounts of urban informal-sector SME dynamics in Zambia [18]; (2) limited access to formal advisory and trade-support services [19]; (3) significant information asymmetry between founders and potential advisors; and (4) a regulatory landscape that includes the Zambia Data Protection Act No. 3 of 2021, which imposes compliance obligations that under-resourced SMEs are ill-equipped to navigate without specialist guidance [20]. This context provides an ideal

setting for examining the application of DSR in resource-constrained environments, as it presents the full range of challenges typical of developing economy contexts.

III.METHODOLOGICAL FRAMEWORK

1. Methodological Overview

Our methodological framework extends the standard DSR model [1] by incorporating adaptations specific to resource-constrained settings. The framework comprises five components:

- **Relevance Cycle Adaptation**: Grounding requirements through purposive expert sampling and sectoral diversity
- **Design Cycle Adaptation**: Iterative development with transparency-oriented design choices (MCDA over machine learning)
- **Rigor Cycle Adaptation**: Integrated theoretical evaluation using multiple frameworks (TTF, UTAUT, D&M)
- **Mixed-Methods Integration**: Sequential exploratory design with qualitative interviews informing quantitative instrumentation
- **Contextual Sensitivity**: Addressing trust deficits, infrastructure constraints, and regulatory complexity

2. Relevance Cycle: Grounding Requirements

In well-resourced settings, requirements gathering typically involves large-scale surveys, extensive user testing, and access to diverse participant pools. In resource-constrained contexts, these approaches are often impractical. Our framework addresses this through:

- **Purposive Expert Sampling:** Rather than attempting statistical representativeness, we employ purposive sampling across diverse sectors to ensure informational depth and theoretical relevance.
- **Semi-Structured Expert Interviews:** We use semi-structured interviews rather than structured surveys to allow participants to raise contextually relevant issues that researchers might not anticipate.
- **Flexible Data Capture:** Participants were offered a choice between audio-recording and non-recorded interviews captured through contemporaneous field notes, in recognition of confidentiality concerns common in advisory and business contexts.
- **Thematic Analysis for Design Requirements:** Thematic analysis of interviews informs both the artifact's feature set and its evaluation criteria.

3. Design Cycle: Iterative Development

The design cycle in resource-constrained contexts requires trade-offs that differ from well-resourced settings. Our framework includes:

- **Transparency-Over-Adaptivity:** We deliberately chose Weighted Multi-Criteria Decision Analysis (MCDA) over machine learning approaches. While MCDA may be less adaptive, it produces transparent, explainable rankings—critical in contexts where digital trust is still developing.
- **Open-Source Technology Stack:** Using standard, open-source frameworks ensures the platform can be maintained without specialist vendor support—essential in resource-constrained settings where ongoing commercial licensing may be unaffordable.

- **Controlled Evaluation Environment:** Early-stage DSR evaluation can be conducted in a controlled, scenario-based environment rather than requiring full operational deployment. This is appropriate for early-stage DSR artifact evaluation and captures user perceptions without the risks and costs of live deployment.

4. Rigor Cycle: Theoretical Evaluation

Standard DSR rigor typically involves evaluation against established theoretical constructs. In resource-constrained settings, we recommend:

- **Parallel Application of Multiple Frameworks:** Rather than testing a single theoretical lens, we apply three complementary frameworks in parallel: Task-Technology Fit (TTF), UTAUT, and the DeLone and McLean IS Success Model. This provides richer evaluation than any single framework alone and reveals context-specific design priorities.
- **Scenario-Based, Prototype-Level Evaluation:** For early-stage artifacts, scenario-based evaluation can yield theoretically meaningful results without requiring full operational deployment. Participants interact with the prototype through guided scenarios before completing evaluation instruments.
- **Effect Size and Power Reporting:** Given the modest sample sizes often feasible in resource-constrained settings, we emphasize reporting effect sizes (f^2) and confidence intervals alongside significance tests, enabling readers to assess practical significance independent of statistical significance.

A notational clarification is necessary here, since three unrelated quantities in this paper are conventionally denoted with the same Greek letters. Sample size ($N = 200$) was planned using an a priori power analysis with a Type I error rate (α) of 0.05 and statistical power ($1-\beta$) of 0.80. α and β strictly refer to power-analysis parameters. They are distinct from (i) Cronbach's alpha (α) internal-consistency coefficients reported per construct and (ii) the standardized regression coefficients (β) reported per predictor. We use the fuller notation "Cronbach's α" and "regression β" throughout the results to avoid ambiguity.

5. Mixed-Methods Integration

Our framework employs a sequential exploratory mixed-methods design [21] that serves two integration functions:

- **Qualitative Phase Informs Prototype Requirements:** Thematic analysis of qualitative interviews directly shapes the prototype's requirements.
- **Qualitative Phase Informs Quantitative Instrument:** The same analysis informs the item pool and construct emphasis of the subsequent quantitative questionnaire.

This integration ensures that the artifact is built around the actual informational environment of its intended users rather than assumptions suited to formalized markets.

6. Contextual Sensitivity

Our framework explicitly addresses three contextual challenges:

- **Trust Deficits:** In low-trust markets, the artifact must function as a verification engine rather than a simple directory [13]. Design choices (transparent MCDA, explainable recommendations) must prioritize trust-building.
- **Infrastructure Constraints:** Designs must account for variable internet connectivity and mobile data costs. The framework includes UI design considerations for low-bandwidth environments.

- **Regulatory Complexity:** The design must address local regulatory requirements (e.g., Zambia Data Protection Act No. 3 of 2021) while helping users navigate compliance obligations.

IV.CASE STUDY: LUMIKIZA APPLICATION

1. Research Context

Zambia lacks a structured digital mechanism for SME-advisor matching. Lumikiza aimed to design, build, and evaluate a prototype addressing this gap. Figure 1 summarizes how the DSR cycles, the integrated theoretical framework, and the Lumikiza artifact relate to one another and to the study's academic and practical contributions.

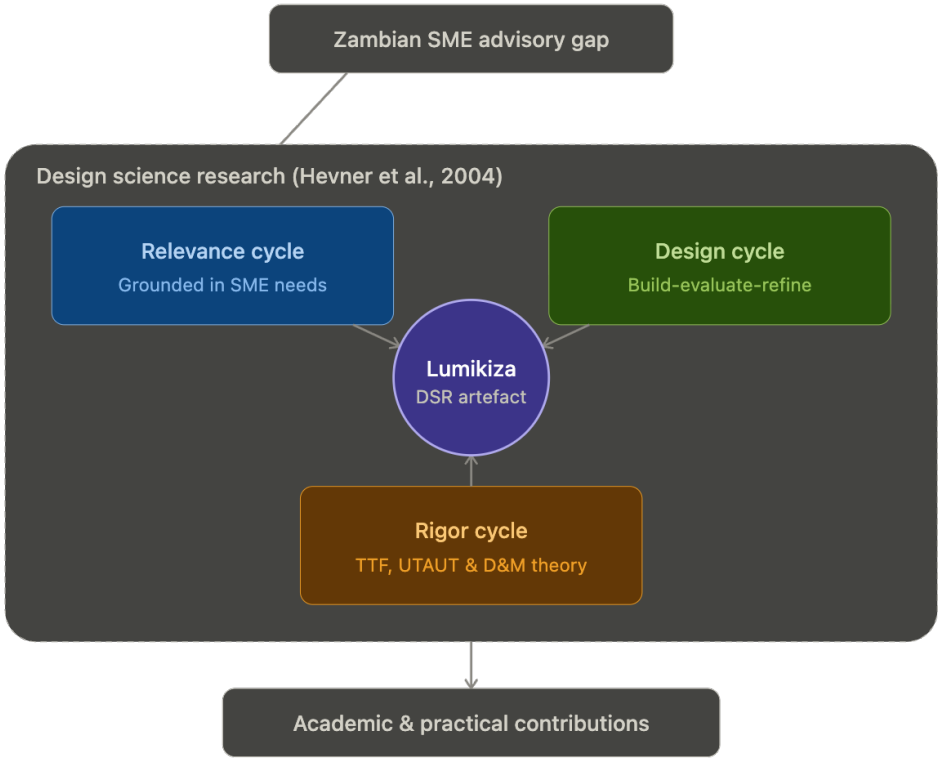


Figure 1. Lumikiza's integrated conceptual framework for digital platform development in a resource-constrained DSR context.

The research was conducted in Lusaka and Livingstone, Zambia, during March-April 2026. These cities were selected because: (1) they have the highest concentration of formally registered SMEs, (2) they represent diverse business models (Lusaka: urban commercial hub; Livingstone: tourism-focused), and (3) both have adequate digital connectivity for platform evaluation

2. Study Site, Population, and Sampling

Qualitative phase (expert advisors): The population was professional advisors and domain experts working with Zambian SMEs. A purposive sample of n = 10 was recruited to ensure sectoral diversity and depth rather than statistical representativeness, consistent with the exploratory, design-oriented nature of DSR evaluation. Inclusion criteria were a minimum of 5 years of professional experience and active engagement in SME advisory work. Participants were purposively drawn across six sectors: Agribusiness, Manufacturing, Mining, Fintech, Retail, and Services.

Quantitative phase (SME founders): The population was SME founders and key decision-makers operating in Lusaka and Livingstone. A purposive sample of N = 200 was recruited, with inclusion criteria requiring that the business have been registered and operating for at least 1 year and that the respondent be involved in strategic decision-making. The sample size was determined using an a priori power analysis for the planned regression analysis ($\alpha = 0.05$, power = $1- \beta = 0.80$).

3. Data Collection, Instrumentation, and Ethics

Qualitative interviews: Semi-structured interviews were conducted with the ten SME advisors and domain experts (coded AD-01 to AD-10). One interview (AD-01) was audio-recorded with explicit consent and transcribed verbatim; the remaining nine participants preferred non-recorded interviews, and their contributions were captured through contemporaneous field notes, which were expanded into thematic memos for coding. An interview guide was used to ensure consistency across interviews while allowing participants to raise contextually relevant issues.

Quantitative instrument: A structured questionnaire was administered to 200 SME founders using a five-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Items were mapped to eight theoretical constructs drawn from TTF [22], UTAUT [23], and the DeLone and McLean model [24]: System Quality, Information Quality, Task-Technology Fit, Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, and Net Benefits. All 200 participants completed the full questionnaire (100% response rate); therefore, no missing-data procedures were required.

Validity and reliability: Content validity for both instruments was established through expert review before data collection. Construct validity was addressed by adapting measurement items from previously validated IS instruments [22], [23], [24]. Internal consistency of the quantitative scales was assessed using Cronbach's alpha, reported by construct. For the qualitative data, trustworthiness was addressed through credibility (member checking where appropriate), dependability (a maintained interview protocol), and confirmability (documented coding decisions).

Ethics: Ethical clearance was obtained from The Copperbelt University before data collection. Participation was voluntary; all participants received a written information sheet and gave written informed consent before any interview was recorded or survey completed. Responses were anonymized using participant codes (SME01–SME200; EXP01–EXP10/AD-01–AD-10), and no identifying information appears in any research output. All data were stored on password-protected devices accessible only to the researcher and supervisor, in compliance with the Zambia Data Protection Act No. 3 of 2021.

4. Application of the Framework

Relevance Cycle

Analysis of the ten expert interviews revealed:

- Trust is the absolute prerequisite for digital adoption
- Word-of-mouth serves as the only current verification system
- A "Diagnostic Gap" prevents SMEs from accessing high-quality information
- Sector-specific expertise is critical
- Cost and credibility concerns are the primary barriers

These findings translated into concrete design requirements: a Trust Assurance score in the matching algorithm, a pre-matching diagnostic survey for SMEs, certification filters and "Verified" expert badges, fee-transparency indicators, and an explicit distinction between "Strategic" and "Technical" experts.

Design Cycle

The Lumikiza prototype was implemented as a modular, decoupled web-based platform as outlined in Table 1 below.

Table 1. Lumikiza Platform Architecture

Layer	Technology
Frontend	Vite, React, TypeScript, Tailwind CSS, shadcn-ui
Application logic	Modular service architecture
Database	PostgreSQL (relational)
Matching engine	Weighted Multi-Criteria Decision Analysis (MCDA)

Matching Engine

The MCDA engine scores registered advisors against five criteria for each advisory request: domain relevance, years of professional experience, sector familiarity, regulatory knowledge, and availability, and returns a ranked list of the top matches.

Each of the five criteria is assigned a weight reflecting its relative importance, informed qualitatively by the thematic emphasis participants placed on domain relevance and sector familiarity during the relevance-cycle interviews. We are explicit about a limitation here: the current implementation uses a fixed, qualitatively-derived weighting scheme. A formal normalization procedure, an explicit aggregation equation, and a quantitative sensitivity analysis of the weights were not conducted at this stage of the evaluation. This was a deliberate scope decision—our priority was demonstrating that expert-interview-derived weighting is feasible and produces explainable rankings—but it is also a genuine limitation of the current prototype. We identify formal weight elicitation (e.g., pairwise comparison or conjoint methods) and sensitivity analysis as the necessary next steps before the matching engine can be considered validated.

MCDA was chosen deliberately over machine learning for its transparent, explainable rankings [25]. One design-evaluation iteration was completed before formal testing, consistent with early-stage DSR evaluation practice [6]; we return to the limits this places on our iterative-DSR claim made above.

Rigor Cycle

The evaluation used an integrated framework combining TTF [22], UTAUT [23], and DeLone and McLean [24], applied as parallel predictors of Net Benefits rather than as a mediated or sequential model (Figure 2). Each evaluation scenario was designed to operationalize specific theoretical constructs, from advisory-request submission (TTF) through recommendation interpretation (information quality, net benefits) to a global post-task assessment (social influence, facilitating conditions, net benefits), consistent with meta-analytic evidence on the determinants of task-technology fit more broadly [26].

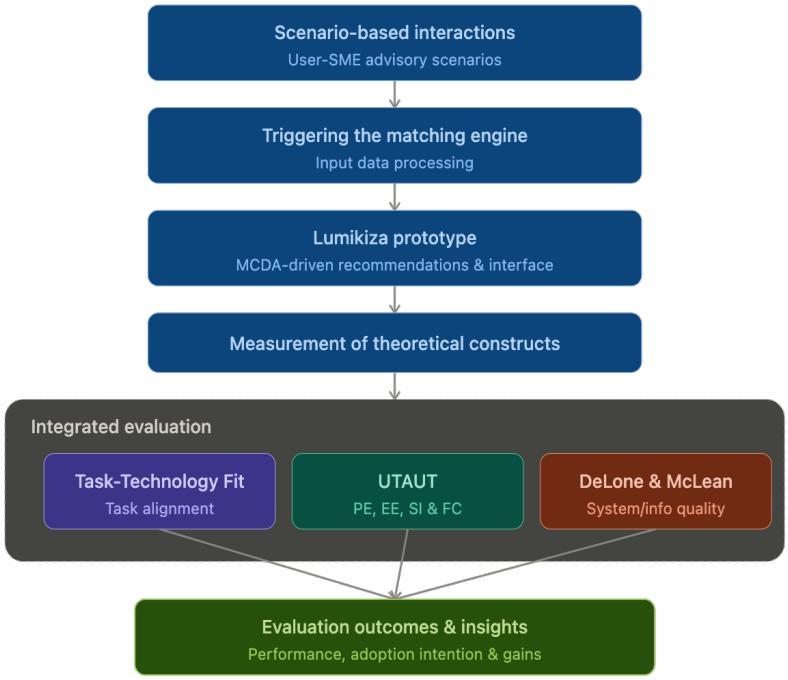


Figure 2. Prototype evaluation and procedural flow, mapping scenario-based interactions to the integrated TTF/UTAUT/D&M measurement framework.

Mixed-Methods Integration

The sequential exploratory design served two integration functions: qualitative interviews shaped the Lumikiza prototype's requirements (particularly the weighting given to regulatory knowledge and sector familiarity), and the same analysis informed the item pool and construct emphasis of the quantitative questionnaire. This integration ensured the prototype was built around the actual informational environment of its users.

Methodological Contributions from the Case

- **Framework for MCDA Weighting:** In the absence of formal ranking exercises, MCDA weights can be derived from thematic analysis of qualitative interviews, with themes that emerge most consistently given greater emphasis—while acknowledging, as above, that this substitutes for rather than replaces formal weight elicitation.
- **Scenario-Based Evaluation:** For early-stage DSR artifacts, scenario-based evaluation can yield theoretically meaningful results without requiring full operational deployment.
- **Integrated Multi-Theoretic Evaluation:** Parallel application of TTF, UTAUT, and DeLone and McLean provides richer evaluation than any single framework alone and can reveal context-specific design priorities.
- **Effect Size Reporting:** In resource-constrained settings with modest samples, reporting effect sizes (f^2) alongside statistical significance enables assessment of practical significance independent of sample size.

V.RESULTS

This section reports the empirical results underlying the evaluation summarized in Section IV. It is presented separately from the Discussion (Section VI) so that findings and their interpretation remain clearly distinguishable.

1. Sample Characteristics

Table 2 summarizes the demographic and firmographic profiles of the 200 SME founders who completed the quantitative questionnaire. The sample skews male (75.0%) and highly educated (94.0% hold at least a Bachelor's degree), and is concentrated in the 35–44 age range and in the Agriculture, Retail, and Services sectors. This profile should be borne in mind when interpreting the generalizability of the findings.

Table 2. Sample Characteristics (N = 200)

Characteristic	Category	n (%)
Gender	Male	150 (75.0%)
	Female	50 (25.0%)
Age	35–39 years	65 (32.5%)
	40–44 years	39 (19.5%)
	Other brackets	96 (48.0%)
Education	Postgraduate	102 (51.0%)
	Bachelor's degree	86 (43.0%)
	Diploma	12 (6.0%)
Sector	Agriculture	64 (32.0%)
	Retail	45 (22.5%)
	Services	44 (22.0%)
	Other	36 (18.0%)
	Manufacturing	11 (5.5%)

2. Reliability and Validity

Internal consistency for each of the eight theoretical constructs was assessed using Cronbach's alpha (Table 3). All constructs exceeded the generally accepted 0.70 threshold except Facilitating Conditions ($\alpha = 0.688$). Because the Facilitating Conditions scale comprised only three items, and Cronbach's alpha is known to be attenuated in short scales [27], an alpha in the 0.65–0.70 range was treated as acceptable for this exploratory evaluation; item-total correlations for all three items exceeded 0.45.

Table 3. Internal Reliability (Cronbach's α) by Construct

Construct	Items	Cronbach's α
Task-Technology Fit (TTF)	7	0.812
Performance Expectancy (PE)	5	0.784
Effort Expectancy (EE)	5	0.821
Social Influence (SI)	3	0.745
Facilitating Conditions (FC)	3	0.688
System Quality (SQ)	6	0.712

Construct	Items	Cronbach's α
Information Quality (IQ)	6	0.756
Net Benefits (NB)	6	0.793

3. Descriptive Construct Scores

All eight constructs achieved mean scores above the 3.0 neutral midpoint of the five-point scale. Effort Expectancy (M = 4.14) and Facilitating Conditions (M = 4.11) scored highest; System Quality (M = 3.71) scored lowest. Task-Technology Fit averaged M = 3.79 across the sample. Standard deviations were tightly bound across constructs, indicating a reasonable degree of consensus among the 200 participants. Inter-construct correlations were consistently positive, and were notably stronger between Net Benefits and both Performance Expectancy and Information Quality than between Net Benefits and the remaining constructs.

4. Predicting Net Benefits: Regression Results

An Ordinary Least Squares (OLS) multiple regression was conducted to predict Net Benefits from the TTF, UTAUT, and D&M constructs. The overall model was not statistically significant ($F = 1.95$, $p = 0.064$) and explained only 6.6% of the variance in Net Benefits ($R^2 = 0.066$; adjusted $R^2 = 0.032$). Within this weak and non-significant overall model, Information Quality ($\beta = 0.167$, $p = 0.038$) was the only predictor to reach the 0.05 significance threshold; System Quality ($p = 0.720$) and Effort Expectancy ($p = 0.534$) were not significant. Table 3 reports the coefficients available for the six predictors captured in the regression output.

Table 4. OLS Regression Results (Dependent Variable: Net Benefits)

Variable	β	SE	95% CI	t	p
(Constant)	2.352	0.520	[1.326, 3.378]	4.525	<.001
Task-Technology Fit	0.113	0.066	[−0.017, 0.243]	1.700	.091
Performance Expectancy	0.073	0.072	[−0.069, 0.215]	1.011	.313
Effort Expectancy	0.035	0.057	[−0.077, 0.147]	0.623	.534
Social Influence	0.036	0.057	[−0.076, 0.148]	0.620	.536
Facilitating Conditions	−0.040	0.055	[−0.148, 0.068]	−0.727	.468
System Quality	0.026	0.074	[−0.120, 0.172]	0.359	.720
Information Quality	0.167	0.080	[0.009, 0.325]	2.085	.038

[a] Note: 95% CIs computed as $\beta \pm 1.972 \times SE$ (t-critical, df = 192). Model fit: $R^2 = 0.0664$, adjusted $R^2 = 0.0324$, $F = 1.9523$ ($p = 0.0636$).

The effect size for the overall model, $f^2 = R^2/(1-R^2) = 0.066/(1-0.066) = 0.071$, exceeds Cohen's [28] convention for a small effect ($f^2 = 0.02$). This indicates a non-trivial effect, but given the model's overall non-significance, we treat both the effect size and the Information Quality coefficient as a preliminary signal rather than a confirmed predictive relationship.

5. Subgroup Analysis: Gender Invariance

A Welch's t-test was used to compare male and female respondents across all eight constructs (Table 4). No statistically significant differences were found (all $p > 0.05$), suggesting that perceived usefulness, effort expectancy, and information quality did not differ by gender in this sample.

Table 5. Independent-Samples t-Test Results (Male vs. Female)

Construct	Male M	Female M	t	p	Sig.
Effort Expectancy	4.147	4.128	0.268	.789	No
Facilitating Conditions	4.111	4.093	0.245	.807	No
Social Influence	3.969	3.980	−0.164	.870	No
Net Benefits	3.907	3.963	−1.029	.306	No
Performance Expectancy	3.899	3.884	0.267	.790	No
Information Quality	3.811	3.803	0.147	.883	No
Task-Technology Fit	3.788	3.814	−0.449	.654	No
System Quality	3.696	3.753	−1.154	.251	No

6. Information Asymmetry Reduction

Before prototype exposure, 76% of participants reported being unable to confidently assess advisors' expertise before first contact, 68% did not know whether advisors understood local regulations, and 62% lacked information about advisors' sector experience. After prototype evaluation, participants were asked whether the platform reduced their uncertainty about advisor fit (Table 5).

Table 6. Response to “Does this platform reduce your uncertainty about advisor fit?”

Response	Frequency	Percentage
Strongly Agree (5)	115	57.5%
Agree (4)	58	29.0%
Neutral (3)	15	7.5%
Disagree (2)	8	4.0%
Strongly Disagree (1)	4	2.0%

Overall, 87% of participants (n = 173) responded positively (4–5 rating), with a mean response of 4.05 (SD = 0.72). This was corroborated qualitatively: Expert AD-04 observed that “*the matching criteria clearly show why each advisor was recommended. This transparency builds confidence in the match*.”

VI.DISCUSSION

1. Key Methodological Insights

Preliminary Signal on Information Quality

Information Quality was the only construct to reach the conventional significance threshold as a predictor of Net Benefits. We emphasize that the overall regression model was not statistically significant ($F = 1.95$, $p = 0.064$) and explained only 6.6% of the variance. We therefore read the Information Quality coefficient as a preliminary, single-sample signal—not as evidence that information quality is established as the dominant driver of platform value in resource-constrained settings generally.

Methodological Implication: Evaluation frameworks for early-stage DSR artifacts should include information quality as a candidate construct, but claims based on an overall model that is not significant should be flagged as preliminary rather than presented as confirmed findings.

Expert Interviews as a Design Tool

Thematic analysis of expert interviews informed both artifact requirements and evaluation instrumentation, supporting the use of small-n purposive sampling for requirements elicitation in resource-constrained settings where large-scale user testing is impractical.

Methodological Implication: Purposive expert sampling combined with thematic analysis can effectively ground DSR requirements when representative sampling is infeasible.

Scenario-Based Evaluation

Scenario-based, prototype-level evaluation yielded theoretically interpretable results without requiring full operational deployment, which is particularly relevant for researchers in resource-constrained settings where live deployment during the research phase is often impractical.

Methodological Implication: Early-stage DSR artifacts can be evaluated in controlled, scenario-based environments, though this comes at the expense of behavioral realism relative to authentic use.

Integrated Theoretical Evaluation

The parallel application of three complementary evaluation frameworks (TTF, UTAUT, D&M) revealed that task alignment and adoption readiness both scored strongly. At the same time, the overall predictive power of the regression model remained weak—a pattern consistent with concerns raised elsewhere about the risk of single-theory technology-adoption studies in MSME contexts [29].

Methodological Implication: Integrated, multi-theoretic evaluation approaches can reveal where constructs diverge, which is analytically useful even when the combined predictive model is not itself significant.

2. Methodological Framework Proposal

Based on our experience, we propose the following framework for DSR in resource-constrained contexts, offered as propositions for testing in other settings rather than as settled guidance:

Table 7. Framework for DSR in Resource-Constrained Contexts

DSR Component	Standard Approach	Adapted Approach
Relevance Cycle	Large-scale surveys, extensive user testing	Purposive expert sampling, semi-structured interviews, thematic analysis
Design Cycle	Adaptive algorithms, proprietary/commercial technology stack	Transparency-oriented algorithms (MCDA), open-source technology stack

DSR Component	Standard Approach	Adapted Approach
Rigor Cycle	Single theoretical framework, full-scale deployment	Integrated theoretical framework, scenario-based evaluation
Evaluation	Large sample sizes for statistical power	Effect size reporting alongside significance tests
Ethics	Standard consent procedures	Address trust deficits explicitly; document consent through accessible mechanisms.
Context	Generic design considerations	Explicitly address trust deficits, infrastructure constraints, and regulatory complexity.

3. Limitations and Considerations

Several limitations bound the interpretation of these findings.

- Single case and context: The methodological insights derive from a single case study in one developing-economy context (Zambia). Generalizability to other resource-constrained settings has not been tested.
- We are explicit about a limitation here: the current implementation uses a fixed, qualitatively derived weighting scheme. A formal normalization procedure, an explicit aggregation equation, and a quantitative sensitivity analysis of the weights were not conducted at this stage of the evaluation. This was a deliberate scope decision. The priority was to demonstrate that expert-interview-derived weighting is feasible and produces explainable rankings. We identify formal weight elicitation (e.g., pairwise comparison or conjoint methods) and sensitivity analysis as the necessary next steps for future iterations.
- One design-evaluation iteration: Only a single build-evaluate-refine cycle was completed before formal testing. This is consistent with early-stage DSR practice, but it limits how strongly we can claim that the iterative character of DSR was demonstrated, rather than asserted.
- Non-significant, low-power regression model: The overall regression model was not statistically significant ($R^2 = 0.066$; $F = 1.95$, $p = 0.064$). The one significant predictor (Information Quality, $\beta = 0.167$, $p = 0.038$) should be interpreted as preliminary and warrants replication with a larger sample before being treated as a settled finding. The "Utilitarian Priority Pattern" language we use is a working label for this preliminary result, not a claim of a generalizable property of SME technology adoption in Zambia or Sub-Saharan Africa.
- Sample composition: The evaluation was conducted in a controlled, scenario-based environment with an educated, urban/peri-urban sample (94% held at least a Bachelor's degree). Findings should not be assumed to generalize to SME founders with only secondary or vocational qualifications, or to rural populations.

4. Implications for Practice

For researchers applying DSR in resource-constrained contexts, we recommend:

- Prioritize transparency over adaptivity in algorithm design to build trust, while budgeting time to formalize and validate any qualitatively derived weighting scheme
- Use expert interviews for requirements elicitation when large-scale user testing is impractical
- Adopt integrated, multi-theoretic evaluation frameworks, but report overall model significance alongside individual predictors so readers can weigh both
- Report effect sizes alongside significance tests, and state plainly when a model is not statistically significant
- Design for low-bandwidth environments and explicitly address trust deficits

- Plan for regulatory compliance as a core design consideration, not an afterthought

VII.CONCLUSION

This paper presented a methodological framework for applying Design Science Research in resource-constrained developing-economy contexts, demonstrated through the Lumikiza digital advisory-matching platform for Zambian SMEs. We showed how the relevance, design, and rigor cycles were operationalized under limited digital infrastructure, trust deficits, and regulatory complexity, and we reported the evaluation results, including a non-significant overall regression model and a single preliminary significant predictor (Information Quality), transparently rather than overstating them.

The framework's five adaptations (Table 6) and the case's methodological lessons are offered as propositions for other researchers to test in resource-constrained contexts, not as generalized findings. Future work should prioritize: replicating the evaluation with a larger and more educationally diverse sample; formally normalizing and sensitivity-testing the MCDA weighting scheme; completing further design-evaluation iterations; and testing the framework's transferability in other developing-economy settings. With these next steps in place, the approach outlined here may support digital platform development work relevant to SDG 8 (Decent Work and Economic Growth) and SDG 9 (Industry, Innovation and Infrastructure).

AUTHOR CONTRIBUTION

Morley Mujansi: Conceptualization, methodology, investigation, data collection, writing – original draft, project administration. Alice P.S. Shemi: Literature review, methodology, supervision, writing – review & editing. Jeremiah J Mwiinga: Data analysis, writing – review & editing. Lilian Mungulube: review & editing.

REFERENCES

[1] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–106, Mar. 2004, doi: 10.2307/25148625.

[2] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, Jun. 2013, doi: 10.25300/misq/2013/37.2.01.

[3] R. Baskerville, A. Baiyere, S. Gregor, A. R. Hevner, and M. Rossi, "Design Science Research Contributions: Finding a Balance between Artifact and Theory," *Journal of the Association for Information Systems*, vol. 19, no. 5, pp. 358–376, 2018.

[4] S. Gebeyehu and H. Twinomurini, "A Collaborative Consumption Digital Platform for Government Organizations using Design Science," *Digital Government Research and Practice*, vol. 3, no. 1, pp. 1–14, Jan. 2022, doi: 10.1145/3510615.

[5] W. C. Gomera and N. R. Mramba, "Deploying design science research in sparse resource settings: Some lessons from design projects in Tanzania," *African Journal of Science Technology Innovation and Development*, vol. 15, no. 6, pp. 705–719, Apr. 2023, doi: 10.1080/20421338.2023.2178786.

[6] J. Müller, S. Würth, T. Schäffer, and C. Leyh, "Toward a framework for determining methods of evaluation in design science research," *Annals of Computer Science and Information Systems*, vol. 39, pp. 231–236, Oct. 2024, doi: 10.15439/2024f7208.

[7] M. Huseynli, U. Bub, and M. C. Ogbuachi, "Development of a method for the engineering of digital innovation using design science research," *Information*, vol. 13, no. 12, p. 573, Dec. 2022, doi: 10.3390/info13120573.

[8] A. Bon, H. Akkermans, and J. Gordijn, "Developing ICT Services in a Low-Resource Development Context," *Complex Systems Informatics and Modeling Quarterly*, no. 9, pp. 84–109, Dec. 2016, doi: 10.7250/csimq. 2016-9.05.

[9] Y. D. Guracho, S. J. Thomas, N. Almutairi, and K. T. Win, "Design and development of a mobile mental health application for individuals with depression and anxiety: design science research methods," *Behavior and Information Technology*, vol. 45, no. 8, pp. 1530–1545, Mar. 2025, doi: 10.1080/0144929x.2025.2481639.

[10] J. Oredo et al., "The Design, Development, and Implementation of a Digital Platform to Enhance Value Co-Creation within an Entrepreneurship Ecosystem in Sub-Saharan Africa," in *WORLD SCIENTIFIC eBooks*, World Scientific, 2024, pp. 13–54, doi: 10.1142/9789819800520_0002.

[11] S. Hlongwane and S. S. Grobbelaar, "A Practical Framework for Value Creation in Health Information Systems From an Ecosystem Perspective: Evaluated in the South African Context," *Frontiers in Psychology*, vol. 13, p. 637883, Jun. 2022, doi: 10.3389/fpsyg. 2022.637883.

[12] S. C. O. Noutchie, "Unlocking SME potential through digital platforms: insights from three African digital economies," *International Journal of Business Ecosystem and Strategy*, vol. 7, no. 4, pp. 106–112, Aug. 2025, doi: 10.36096/ijbes.v7i4.859.

[13] J. Mou, D. Shin, and J. Cohen, "Understanding trust and perceived usefulness in the consumer acceptance of an e-service: a longitudinal investigation," *Behavior and Information Technology*, vol. 36, no. 2, pp. 125–139, Jul. 2016, doi: 10.1080/0144929x.2016.1203024.

[14] M. Bode, L. Küng, and W. Brenner, "Characterizing the digital transformation of IT consulting services," *IET Software*, vol. 16, no. 6, pp. 469–484, 2022.

[15] World Bank, "Small and medium enterprises (SMEs) finance," World Bank Group, 2022. [Online]. Available: https://www.worldbank.org/en/topic/semifinance

[16] L. M. Lekhanya, "Leadership and corporate governance of small and medium enterprises (SMEs) in South Africa: Public perceptions," *Corporate Ownership and Control*, vol. 12, no. 3, pp. 215–222, 2015.

[17] E. Endris and A. Kassegn, "The role of micro, small and medium enterprises (MSMEs) to the sustainable development of sub-Saharan Africa and its challenges: a systematic review of evidence from Ethiopia," *Journal of Innovation and Entrepreneurship*, vol. 11, no. 1, p. 20, Feb. 2022, doi: 10.1186/s13731-022-00221-8.

[18] A. Mubita, M. Munalula, M. Libati, N. S. Mwale, and V. Kanyamuna, "Urban Informality and Small Scale Enterprise (SME) Development in Zambia: An Exploration of Theory and Practice," *Journal of Behavioral Economics, Finance, Entrepreneurship, Accounting and Transport*, vol. 5, no. 1, pp. 19–29, 2017.

[19] N. Mwanza and P. Silwimba, "Investigating the levels of SMEs' participation in international trade in Zambia: A case study of SMES in Lusaka's Kamwala Trading Center," *Zenodo*, Dec. 2025, doi: 10.5281/zenodo. 17922205.

[20] [Government of Zambia, *The Data Protection Act No. 3 of 2021*. Government Printers, 2021.

[21] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*, 3rd ed. SAGE Publications, 2018.

[22] D. L. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," *MIS Quarterly*, vol. 19, no. 2, pp. 213–236, 1995.

[23] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003.

[24] W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: A ten-year update," *Journal of Management Information Systems*, vol. 19, no. 4, pp. 9–30, 2003.

[25] S. Corrente, S. Greco, M. Kadzinski, and R. Slowinski, "Robust ordinal regression in preference learning and ranking," *Machine Learning*, vol. 93, no. 2–3, pp. 381–422, 2013.

[26] S. Chavarnakul, Y.-C. Lin, A. Khan, and S.-C. Chen, "Exploring the determinants and consequences of Task-Technology Fit: A meta-analytic structural equation modeling perspective," *Emerging Science Journal*, 2024.

[27] N. Schmitt, "Uses and abuses of coefficient alpha," *Psychological Assessment*, vol. 8, no. 4, pp. 350–353, 1996.

[28] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.

[29] T. V. R. Mantik, D. D. Kandouw, N. G. Karouwan, and E. R. Simanjuntak, "Self-efficacy and technology adoption for micro, small, and medium enterprises: An integrated model of Task-Technology Fit and Technology Acceptance," *Journal of World Science*, vol. 3, no. 2, pp. 271–287, 2024.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Brain-Computer Interface, Brain Science, and Artificial Intelligence

Stevo Bozinovski
Dept of Computer Science and Mathematics
South Carolina State University
Orangeburg, USA
sbozinovski@scsu.edu
ORCID:0000-0002-4360-431X

Adrijan Božinovski
Faculty of Informatics
American University of Europe FON
Skopje, Macedonia
adrijan.bozinovski@fon.edu.mk
ORCID:0000-0003-2820-4016

Abstract
The paper comments on the Brain-Computer Interface, which relies on brain rhythms and other brain signals. It presents a reasoning in Brain Science, particularly regarding brain waves for the Brain Computer Interface. Also, the paper presents a relation between brain rhythms and choice, as a notion of Intelligence.

Keywords: Brain Computer Interface, Artificial Intelligence, brain rhythms.

I. INTRODUCTION

Brain rhythms are oscillations in a human electroencephalogram (EEG) defined by their frequency ranges. The ranges are defined *empirically*, and their definitions vary from one research report to another. For example, one study may define the alpha rhythm range as 8-13 Hz (oscillations per second), and another study as 8-12 Hz. Brain rhythms are an important area of Brain Computer Interface (BCI) research, i.e., research on directly controlling external objects with EEG signals. This paper first provides a brief review of the origins of Brain-Computer Interface (BCI) research, highlighting the role of the alpha rhythm in the pioneering control of a robot using BCI. The paper seeks a mathematical basis for a theoretically grounded definition of the frequency ranges of brain rhythms.

II. BRAIN-COMPUTER INTERFACE

Research on Brain-Computer Interface (BCI) began with the BCI challenge [1]. The challenge was to find a way to control external objects (outside the human body) using EEG signals. He advised using various EEG signals, including EEG rhythms and event-related potentials. Specifically, he challenged the use of the Contingent Negative Variation (CNV) potential. Also, he mentioned using electrooculography (EOG) signals. The first five applications of BCI, after Vidal's (1973) challenge were: control of a cursor-like object on a screen [2], using the Visual Evoked Potential (VEP); control of a buzzer [3-6] using the CNV potential; control of a physical object (a robot) [7-10] using the alpha rhythm; and writing text on a screen [11] using the P300 potential. Brain rhythms have been used in BCI since the beginning of BCI [8] when the alpha rhythm was used to control a physical object to solve the long-lasting psychokinesis challenge. A review on the Brain Computer Interface is given in [12-15].

III. BRAIN THEORY

Brain Theory seeks to explain phenomena in the brain, including those related to electrical signals. The electrical oscillations in the brain were first observed in animals [16] and later in humans [17]. Berger named it Elektrenkephalogramm, which, in English literature, is translated as electroencephalogram (EEG). He also observed that, in an EEG, a certain frequency rhythm is dominant under certain conditions, and named it the

alpha rhythm. Since then, various other rhythms have been observed and named delta, theta, beta, and gamma, with additional variants. The alpha rhythm has three variants, depending on the location from which it is recorded. If recorded from the sensorimotor area, it is often called the mu rhythm. If recorded from the auditory area, it is often called the tau rhythm. The alpha rhythm from the visual area and all other areas is named the alpha rhythm. Brain signals can be divided into ongoing brain oscillations and transient potentials. Both transient potentials and brain oscillations can be event-related. Oscillations can be event-related because an event, internal or external, can change the amplitude and frequency of a brain oscillation. Later in the text, we will focus more on brain rhythms. Figure 1 shows the obtaining (filtering out) of classical brain rhythms (alpha, beta, delta, and theta) from an EEG signal.

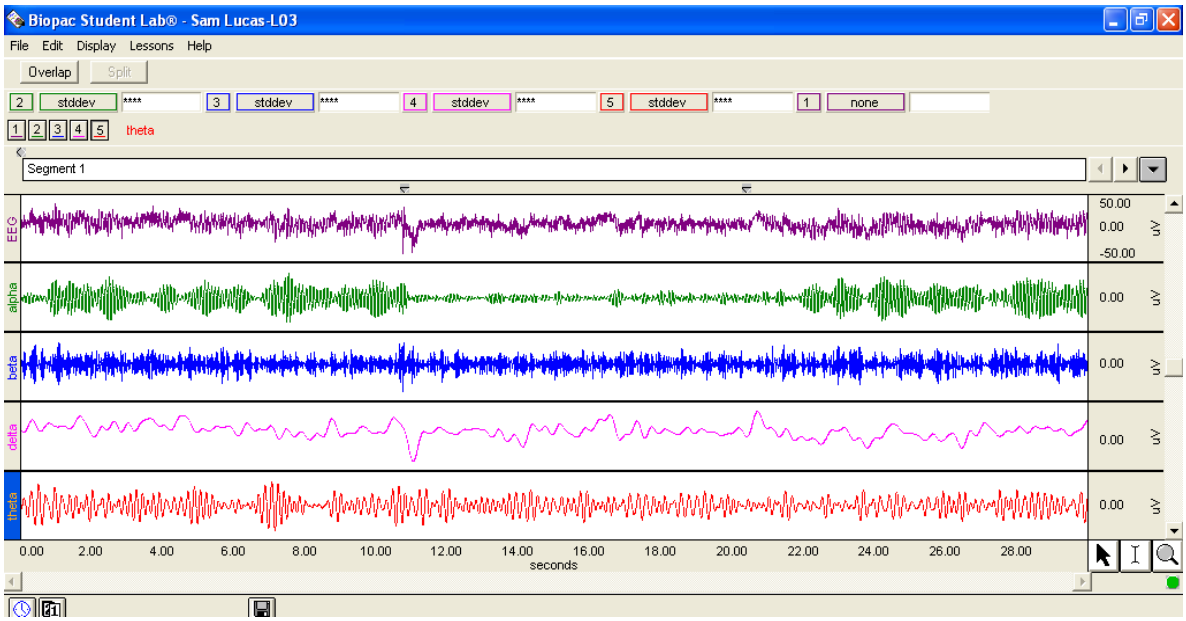


Figure 1. EEG rhythms obtained by Biopac equipment

On top of Figure 1 is the EEG signal obtained from a subject. The unprocessed EEG signal is shown above, whereas the signals below it are obtained after the unprocessed EEG is band-pass filtered. The definitions of the frequency ranges of the filtered signals in Figure 1 are: delta (1-5 Hz), theta (4-8 Hz), alpha (8-13 Hz), and beta (13-30 Hz), all computed using Biopac equipment [18]. There are various interpretations of brain states associated with a particular brain rhythm. According to one interpretation, the alpha rhythm has its highest amplitudes when the visual system is not doing any processing. For example, closing the eyes and not thinking about something will produce high amplitudes in the alpha band. The beta rhythm is dominant in an EEG when the brain is engaged in calculations or image processing. Basically, the appearance of the beta rhythm blocks the alpha rhythm. The border between the alpha and beta can be used as a BCI switch. The first BCI control of a physical object (a robot) used the alpha/beta switch [8].

IV. BRAIN FREQUENCIES AND NUMBER SEQUENCES

The brain rhythms are defined by their frequency bands. The long-lasting issue in Brain Theory is that the frequency bands are defined empirically. As a consequence, in one study, the alpha band is defined as 7-13Hz, in another as 8-12Hz, and in another as 8-14Hz, among other empirical definitions. Thus, a challenge in Brain

Theory and Information Science in general is to find a theoretical process underlying the definition of brain rhythms. Our approach toward brain rhythms is that they are connected to the operation of choice. The intuition is that the brain works on higher frequencies depending on the choice it makes at a particular moment. In the case of two choices (e.g., open-close), the brain operates predominantly within the alpha-frequency range. A question that emerges is: what is a known mathematical concept dealing with choices? It is Pascal's triangle. Pascal's triangle is a pattern for generating integers, level by level, starting with the number 1 at level n=0. It can be constructed using the rule that each number in a level n is the sum of the two numbers above it in level n-1, and each level starts and ends with the number 1. A well-known interpretation of the rows of Pascal's Triangle is that they represent the coefficients of the expansion of (a+b) ^n for n=0, 1, etc.

The known sequences of integers are listed in the Online Encyclopedia of Integer Sequences (OEIS). Examples of sequences in the OEIS are obtained in our previous work [19]. In OEIS, one can find sequences of integers generated by Pascal's triangle. Our contribution is that the central frequencies of the brain rhythms can be defined by the median values of Pascal's Triangle. Figure 2 illustrates our proposed definition of the central frequencies of brain rhythms. The idea is simply that the central frequencies of brain rhythms are the medians of each row of the Pascal triangle. Note that the median in rows with an even number of values is the average of the adjacent values. Each central frequency, along with its range, can be considered a fuzzy number.

n=0											1								
n=1										1		1							
n=2									1		2		1						
n=3								1		3		3		1					
n=4							1		4		6		4		1				
n=5						1		5		10		10		5		1			
n=6					1		6		15		20		15		6		1		
n=7				1		7		21		35		35		21		7		1	
n=8			1		8		28		56		70		56		28		8		1
n=9		1		9		36		84		126		126		84		36		9		1

Figure 2. Pascal's triangle: Our definition of the central frequencies and ranges of brain rhythms

The obtained sequence is 3,6,10,20,35,70,126. With it, we present a mathematical process that defines the central frequencies of brain rhythms. Based on such an analysis, we can write a mathematical formula that defines the frequency bands of brain rhythms.

Let $S(n,k) = 2 \times 1\ matrix, n\ , sonraki\ satır, k\ son\ matris\ , sağ\ parantez\ n = 0,1,\ldots,9;\ k = 0,\ldots,n;\ be\ coefficients\ of\ the\ expansion\ of\ (a+b)\text{^}n$ n = 0,1,...,9; k=0,...,n; be coefficients of the expansion of the $(a+b)^n$.

kkk = k $\binom{n}{M}$ be a median value of $\binom{n}{k}$, k= 0, ..., n.

Then, the brain rhythms' frequency bands are defined by the following mathematical formula:

$$B(n)=\binom{n+2}{M} \pm \binom{n}{M} \quad (1)$$

Over a band B(n), n = 0,...,7, we define a naming function name(B(n)) = (Ldelta, Hdelta, theta, alpha, beta, gamma, gamma2, gamma3), which yields a sequence of names for the brain rhythms.

Equation (1), along with the names, defines the brain rhythms as shown in Table 1.

Table 1. Brain frequency ranges, defined using central frequencies and ranges

Brain Rhythm	Central Frequency and Deviation	Frequency Range
Ldelta	2 ± 1 Hz	1-3 Hz
Hdelta	3 ± 1 Hz	2-4 Hz
theta	6 ± 2 Hz	4-8 Hz
alpha	10 ± 3 Hz	7-13 Hz
beta	20 ± 6 Hz	14-26 Hz
gamma	35 ± 10 Hz	25-45 Hz
gamma2	70 ± 20 Hz	50-90 Hz
gamma3	126 ± 35 Hz	91-161 Hz

A comparison between our theoretically derived ranges and empirical ranges for brain rhythms reported in the literature is shown in Table 2.

Table 2. A comparison of our theoretical ranges of brain rhythms and empirical ranges in the literature

source	year	Ldelta	Hdelta	theta	alpha	beta	gamma	gamma2	gamma3
This paper	2026	1-3	2-4	4-8	7-13	14-26	25-45	50-90	91-161
Berger [17]	1929				8-12	14-30			
Jasper,Andr. [20]	1938						35-45		
IFSECN [21]	1974	0.5-4		4-8	8-12	12-30	>30		
Rice et al. [22]	1989	0-3.5		3.5-7	7-13	13-40			
Buzsaki,Drag.[23]	2004	0.5-1.43	1.5-4	4-10		10-30	30-80		80-200
Hutchison et [24]	2004	0.3-2	2-7		7-13	11-30	30-80		
Kelly et al. [25]	2005				8-14	18-26			
VanRullen,Du[26]	2011	0-4		4-8	8-14	14-30	30-80		
Buzsaki, Wang[27]	2012						30-90		90-150
Jatupaiboon e[28]	2013	1-3		4-7	8-13	14-30	31-50		
Dimitrakopoulo[29]	2017	1-4		4-7.5	8-13	14-26			
Perotti, DeVit[30]	2019			4-12			25-45		
Li et al. [31]	2023			4-7	7-13	13-30			

Therein, in green, are the frequency ranges that exactly match our theoretical ranges to the empirical frequency ranges. In yellow are the frequency ranges that match only one end of our theoretical ranges. For example, the empirical range 18-26 Hz matches only the 26-border of the theoretical range 14-26 Hz. Our theoretically derived central frequencies and ranges of brain rhythms align well with the empirical definitions in the literature [17],[20-31]. So, according to our analysis, the central frequencies are defined from the median values of the rows of Pascal's Triangle. The deviations around the central frequency of a band are also defined with values from Pascal's Triangle. However, while the central frequencies have fixed values and represent a fuzzy number, the definition does not require that the deviations be fixed. There is a possibility that deviations are defined using different values than those in Pascal's Triangle. For example, instead of defining the delta band as 3±1 Hz, it can be defined as 3±2 Hz (range 1-5Hz).

V. ARTIFICIAL INTELLIGENCE AND ENGINEERED PSYCHOLOGY

The Pascal Triangle can be observed as an expansion of $(a+b)^n$, where the coefficients represent the operation of choice. This suggests that the brain rhythms are related to the very important *cognitive operation of choice*. Indeed, the numbers inside Pascal's Triangle represent the number of combinations (choices) when one chooses a k-element subset from a set of n elements. The most commonly used notation is sol parantez 2 denklem denklem dizisi n son denklem, k son denklem dizisi, sağ parantez, which is read as "n choose k". The n-th row of Pascal's Triangle gives the number of choices for each k from 0 to n. We do not claim that the choice defined in the Pascal Triangle is the same as the one defined in cognitive science. However, we found that it is an interesting relation. Artificial Intelligence has many directions, and one of them is Engineered Psychology [32], in which concepts of human psychology, such as motivation, expectation, choice, emotion, and personality, are explored.

VI. CONCLUSION

Starting our discussion with the Brain-Computer Interface, an application of Brain Theory, we focused on brain rhythms and a theory for their definition. We found that the median sequence of Pascal's Triangle can serve as a theoretical basis for the central frequencies and ranges of brain rhythms. Viewing the extension of Artificial Intelligence toward Engineered Psychology, the paper also suggests that the operation of choice, as an important feature of Intelligence, lies in the background of brain rhythms.

REFERENCES

[1] J. Vidal, "Toward direct brain-computer communication," Annual Review of Biophysics and Bioengineering, pp. 157-180, 1973.

[2] J. Vidal, "Real-time detection of brain events in EEG," Proceedings of the IEEE, vol. 65, pp. 633–641,1977.

[3] L. Bozinovska, M. Sestakov, G. Stojanovski, and S. Bozinovski, "Intensity variation of the CNV potential during the biofeedback training guided by a personal computer," [abstract] (In Croatian). Neurologija; suppl. 2, vol. 37, p. 76, 1988

[4] L. Bozinovska, S. Bozinovski, G. Stojanov, and M. Sestakov, "Introduction of biofeedback in the CNV paradigm" (In Croatian), Proc Conference ETAN, Novi Sad, pp. XII.93-98, 1989.

[5] L. Bozinovska, G. Stojanov, M. Sestakov, and S. Bozinovski, "CNV pattern recognition - A step toward cognitive wave observation," In: L. Torres, E. Masgray, M. Lagunas (editors). Signal Processing: Theories and Applications. Proc Fifth European Signal Processing Conference (EUSIPCO 90); Barcelona. Amsterdam: Elsevier Science Publishers, pp. 1659-1662, 1990.

[6] L. Bozinovska, S. Bozinovski, and G. Stojanov, "Electroexpectogram: Experimental design and algorithms" Proc. IEEE International. Biomedical. Engineering Days, Istanbul, pp. 58–60, 1992.

[7] S. Bozinovski, M. Sestakov, and L. Bozinovska, "Control of a mobile robot using alpha rhythm from a human brain," (In Croatian), Proc. Symposium JUREMA, Zagreb, pp. 247-249,1988.

[8] S. Bozinovski, M. Sestakov, and L. Bozinovska, "Using EEG alpha rhythm to control a mobile robot," In: G. Harris, C. Walker (editors) Proc. 10th Annual Conference of the IEEE Engineering in Medicine and Biology Society, track 17, Biorobotics, New Orleans, vol. 10, pp. 1515–1516, 1988.

[9] S. Bozinovski, M. Sestakov, G. Stojanov, and L. Bozinovska, "Bioelectric control of mobile robots" (In Macedonian) Proc. 6th Yugoslav symposium for applied robotics and flexible automation, Novi Sad, pp., 237-242,1989.

[10] S. Bozinovski, “Mobile robot trajectory control: From fixed rails to direct bioelectric control,” In: O. Kaynak (editor) Proc. IEEE International Workshop on Intelligent Motion Control, Istanbul, vol 2, pp. 463-467, 1990.

[11] L. Farwell, and E. Donchin, “Talking off the top of your head: a mental prosthesis utilizing event-related brain potentials,” Electroencephalography and Clinical Neurophysiology, vol. 70, pp. 510–523, 1988.

[12] M. Lebedev “Augmentation of sensorimotor functions with neural prostheses,” Opera Medica and Physiologica, 2 (3):211–227, 2016.

[13] M. Lebedev, and M. Nicolelis, “Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation,” Physiological Review vol. 97, pp. 737-867, 2017.

[14] X. Liu and J. van der Spiegel, Brain Machine Interface, Springer-Verlag, 2017.

[15] S. Bozinovski and L. Bozinovska, “Brain-Computer Interface in Europe: The thirtieth anniversary,” Automatika, vol. 60, no. 1, pp. 36-49, 2019.

[16] V. Pravdich-Nemenskij, “Ein Versuch der Registrierung der elektrischen Gehirnerscheinungen,” Zentralblatt für Physiologie, vol. 27, pp. 951-960, 1913.

[17] H. Berger, “Über das Elektrenkephalogramm des Menschen,” Archiv für Psychiatrie, vol. 87, pp. 527-570, 1929.

[18] Biopac Student Lab. Biopac Systems, Inc., 2013.

[19] A. Bozinovski and G. Tanev, “A277267”, Online Encyclopedia of Integer Sequences, 2016.

[20] H. Jasper, and H. Andrews, "Brain potentials and voluntary muscle activity in man." Journal Neurophysiology1, 87–100, 1938.

[21] M.Steriade, P. Gloor, R. Llinás, F. Lopes da Silva, and M. Mesulam, "Basic mechanisms of cerebral rhythmic activity" Electroencephalogr Clin. Neurophysiol. vol. 76, p. 481–508, 1990.

[22] D. Rice, S. Potkin, Y. Jin, R. Isenhart, C. Heh, J. Sramek, J. Costa, and C. Sandman, "EEG alpha photic driving abnormalities in chronic schizophrenia." Psychiatry Research 30, 313-324, 1989.

[23] G. Buzsaki and A. Draguhn, "Neural oscillations in cortical networks," Science 304, 1926-1929, 2004.

[24] W. Hutchison, J. Dostrovsky, J. Walters, R. Courtemanche, T. Boraud, J. Goldberg, and P. Brown, "Neuronal oscillations in the basal ganglia and movement disorders: Evidence from whole animal and human recordings." The Journal of Neuroscience 24(42),9240 –9243, 2004.

[25] Kelly, S., Lalor, E., Reilly, R., and Foxe, J. “Independent brain-computer interface control using visual spatial attention-dependent modulations of parieto-occipital alpha,” Proc. 2nd International IEEE EMBS Conference on Neural Engineering. Arlington, p. 667-670, 2005.

[26] R. Van Rullen, and J. Duboi, "The psychophysics of brain rhythms," Frontiers in Psychology, Perception Science 2, Article 203, 1-10, 2011.

[27] G. Buzsaki, and X-J. Wang, "Mechanisms of gamma oscillations." Annual Review of Neuroscience 35, 203-225, 2012.

[28] N. Jatupaiboon, S. Pan-ngum, and P. Israsena, "Real-time EEG-based happiness detection system," The Scientific World Journal 2013, Article ID 618649, 12 pages, 2013.

[29] G. Dimitrakopoulos, I. Kakkos, Z. Dai, J. Lim, J. deSouza, A. Bezerianos, and Y. Sun, "Task-independent mental workload classification based upon common multiband EEG cortical connectivity." IEEE Transactions on Neural Systems and Rehabilitation Engineering 25(11), 1940-1949, 2017.

[30] L. Perotti, J. DeVito, D. Bessis, and Y. Dabaghian, “Discrete structure of the brain rhythms,” Nature 9:1105, 2019.

[31] Z. Li, W. Nan, S. Pun, M. Wai, A. Rosa, and F. Wan, "Modulating individual alpha frequency through short-term neurofeedback for cognitive enhancement in healthy young adults," Brain Science 13(6), 926, 2023.

[32] S. Bozinovski and L. Bozinovska, "Beyond artificial intelligence toward engineered psychology," In. M. Uliery, P. Palensky, R. Doursat (eds.) IT Revolutions, Springer, 181-185, 2009.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Scalability Analysis of CNN-Based sEMG Hand Gesture Recognition Across Increasing Gesture Repertoires

Yasir ÖZLÜK
Electrical and Electronics Engineering,
Graduate School of Natural and Applied Sciences,
Gazi University
Ankara, Türkiye
24833001008@gazi.edu.tr
ORCID:000-0002-7812-3516

Eda AKMAN AYDIN
Electrical and Electronics Engineering,
Faculty of Technology,
Gazi University
Ankara, Türkiye
edaakman@gazi.edu.tr
ORCID:0000-0002-9887-3808

Abstract
Surface electromyography (sEMG)-based hand movement recognition systems have been widely employed in various applications, including prosthetic control, rehabilitation, and human–machine interaction. In these applications, evaluating the effects of movement repertoire size and electrode configuration on classification performance is essential to developing systems that reliably recognize a larger number of hand movements. In this study, the scalability of sEMG-based hand movement recognition systems was evaluated by progressively expanding the movement repertoire, and the effects of different electrode configurations were systematically investigated. Experiments were conducted on the GrabMyo dataset, with four movement-repertoire scenarios comprising 4, 8, 12, and 16 movements. For each scenario, the same Convolutional Neural Network (CNN) architecture was evaluated under three different electrode configurations: wrist (12 channels), forearm (16 channels), and combined wrist–forearm (28 channels). The results demonstrated that classification performance decreased across all electrode configurations as the movement repertoire increased. However, the 28-channel electrode configuration consistently achieved the highest classification accuracy across all scenarios and exhibited greater robustness to the increase in movement repertoire size. These findings provide practical insights into selecting an appropriate movement repertoire size and electrode configuration for different sEMG-based hand movement recognition applications.

Keywords: surface electromyography (sEMG), hand movement recognition, movement repertoire size, class scalability, electrode configuration, convolutional neural network (CNN)

I. INTRODUCTION

Electromyography (EMG) is a biophysiological signal obtained by recording the electrical activity generated during voluntary muscle contractions, providing valuable information about muscle function. Surface electromyography (sEMG) is one of the most widely used physiological data sources in hand movement recognition systems because it enables the non-invasive measurement of these electrical signals. sEMG-based hand movement recognition systems have been employed in a wide range of applications, including prosthetic control, rehabilitation, human–computer interaction, virtual and augmented reality, sign language recognition, assistive technologies, and intelligent robotic systems [1-6]. In recent years, advances in sensor technologies, wearable electronic systems, and artificial intelligence-based methods have enabled these systems to move beyond controlled laboratory environments and become increasingly applicable to real-world scenarios. Accordingly, there has been growing interest in and need for scalable sEMG-based hand movement recognition systems that reliably recognize larger movement repertoires while accommodating the diverse requirements of real-world applications.

The movement repertoire is the complete set of distinct hand and forearm movements that the system is intended to discriminate. The movement repertoire size refers to the number of movement classes included in the classification problem. Different sEMG-based applications exhibit substantial differences in the movement repertoires they require. For example, recognizing a limited number of hand movements may be sufficient for basic prosthetic control or simple human–computer interaction applications [7], [8]. In contrast, advanced prosthetic systems, robotic manipulation, sign language recognition, and virtual/augmented reality applications [9], [2] require the reliable classification of substantially larger movement repertoires. However, increasing the movement repertoire size increases the number of classes that must be discriminated, thereby making the classification problem more complex and directly affecting system performance [10]. Therefore, evaluating how classification performance changes with increasing movement repertoire size is important for assessing the scalability of sEMG-based hand movement recognition systems. Nevertheless, most studies in the literature have evaluated performance using a fixed number of movements. In contrast, systematic investigations into the effect of progressively increasing the size of the movement repertoire on classification performance remain limited.

In addition to movement repertoire size, electrode configuration is another important factor influencing the performance of sEMG-based hand movement recognition systems. Electrode placement, the number of channels, and the muscle groups covered are fundamental hardware parameters that directly affect the representation of muscle activity and the information content and discriminative characteristics of the recorded sEMG signals. Various electrode configurations involving the wrist, forearm, or both regions have been proposed in the literature, and their effects on classification performance have been evaluated across different datasets and applications [11], [12]. However, the vast majority of existing studies have used fixed movement repertoires. In contrast, systematic studies of how performance across different electrode configurations changes as the movement repertoire size increases remain limited. This limitation makes it challenging to determine the most appropriate electrode configuration for different application requirements. It highlights an important research gap concerning the scalability of sEMG-based hand movement recognition systems.

In this study, the effects of progressively expanding the movement repertoire and employing different electrode configurations on the scalability of sEMG-based hand movement recognition systems were systematically investigated. To this end, the GrabMyo dataset was employed, and the movement repertoire was progressively expanded from gross motor movements to fine motor movements. Specifically, following forearm rotation and grasping movements, isolated single- and two-finger movements, as well as more complex finger combinations, were progressively incorporated into the repertoire, resulting in four experimental scenarios comprising 4, 8, 12, and 16 movements. This experimental design enabled the evaluation of the increasing complexity of the classification problem associated with an expanding movement repertoire within a controlled experimental framework.

Each experimental scenario was analyzed using the same Convolutional Neural Network (CNN) architecture across three electrode configurations: wrist-only (12 channels), forearm-only (16 channels), and combined wrist and forearm (28 channels). The same network architecture and training strategy were employed across all experiments to ensure that the observed differences in performance were attributable solely to movement repertoire size and electrode configuration. Accordingly, changes in classification performance with increasing movement repertoire, as well as the effects of different electrode configurations on these changes, were directly compared. The study aims to inform design decisions on selecting an appropriate movement repertoire size and electrode configuration for different application requirements.

II. METHODOLOGY

1. Dataset

The GrabMyo dataset was employed in this study because it provides three different channel configurations and includes a diverse range of finger, hand, wrist, and forearm movements. The dataset contains EMG recordings from 43 healthy participants, collected over three separate sessions, comprising 17 classes corresponding to 16 movements and a rest condition. In this study, signals corresponding to the 16 movement classes were used. The data were recorded at a sampling frequency of 2048 Hz using 28 channels, including 12 placed around the wrist and 16 around the forearm [13], [14].

2. The Proposed CNN Model

In this study, a lightweight one-dimensional Convolutional Neural Network (CNN) model was employed to classify sEMG signals. As illustrated in Fig. 1, the proposed architecture consists of three sequentially arranged Conv1D blocks. The first convolutional layer contains 32 filters, whereas the second and third layers contain 64 filters. A kernel size of 7 samples, same padding, and L2 regularization ($\lambda = 1 \times 10^{-5}$) were applied to all convolutional layers. The input to the model consists of segmented sEMG signals with dimensions of (204 x C). Local temporal patterns are extracted through the following one-dimensional convolution operation:

$$y_f[t] = \sum_{c=0}^{C-1} \sum_{k=0}^{K-1} W_f[c,k] \cdot x_c[t+k] + b_f \ , \quad f = 0,1,2,\dots F-1 \tag{1}$$

where (C) denotes the number of input channels, (K) represents the kernel size, and (W) and (b) denote the filter weights and bias parameters, respectively.

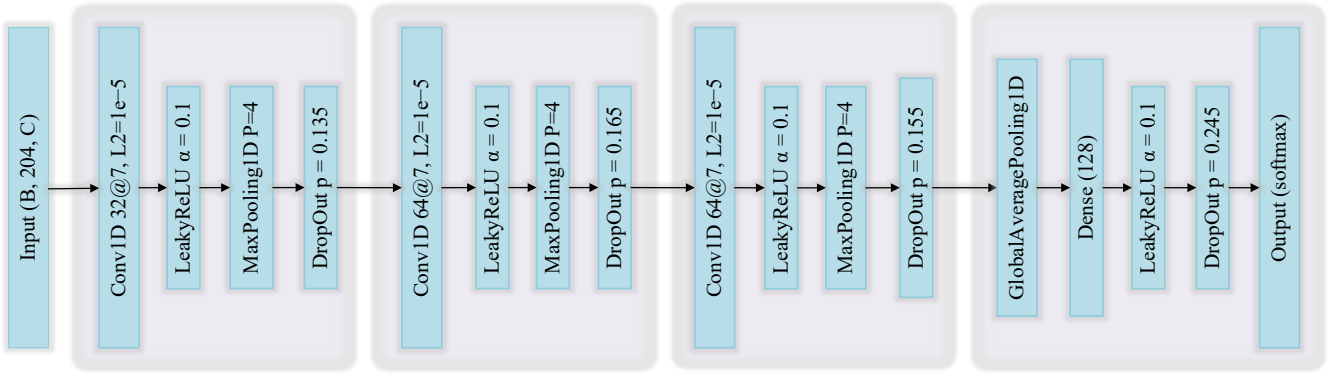


Figure 1. Proposed model flowchart

Each convolutional layer is followed by a LeakyReLU activation function ($\alpha = 0.1$), MaxPooling1D (pool size = 4), and Dropout with rates of 0.135, 0.165, and 0.155, respectively. The feature maps generated by the final convolutional block are summarized using GlobalAveragePooling1D and subsequently processed by a fully connected layer with 128 neurons. The output layer consists of 4–16 neurons with a Softmax activation function, depending on the movement repertoire scenario. The model was trained for a maximum of 100 epochs with a batch size of 64 using the Adam optimization algorithm with a learning rate of (1×10^{-4}) and the categorical cross-entropy loss function. To mitigate overfitting, Early Stopping with a patience of 5 epochs was employed. In addition, when the validation loss failed to improve, the learning rate was dynamically reduced using the ReduceLROnPlateau strategy, with a reduction factor of 0.7, a patience of 3 epochs, and a minimum learning rate of (1×10^{-7}).

3. Expansion of the Movement Repertoire

To evaluate the effect of movement repertoire size on classification performance, the movements in the GrabMyo dataset were grouped into four experimental scenarios comprising 4, 8, 12, and 16 movements, respectively. The movement repertoire was constructed by progressively adding movements from the categories of Gross and Whole-Hand Movements; Forearm Rotation and Grasping Movements; Isolated Single- and Two-Finger Movements; and Complex Finger Combination Movements, thereby representing increasing levels of biomechanical complexity. This progressive design enabled the increasing complexity of the classification problem associated with an expanding movement repertoire to be evaluated within a controlled experimental framework.

Each experimental scenario was analyzed using the same Convolutional Neural Network (CNN) architecture under three different electrode configurations: wrist (12 channels), forearm (16 channels), and combined wrist and forearm (28 channels). The grouping strategy used to construct the movement repertoires is illustrated in Fig. 2.

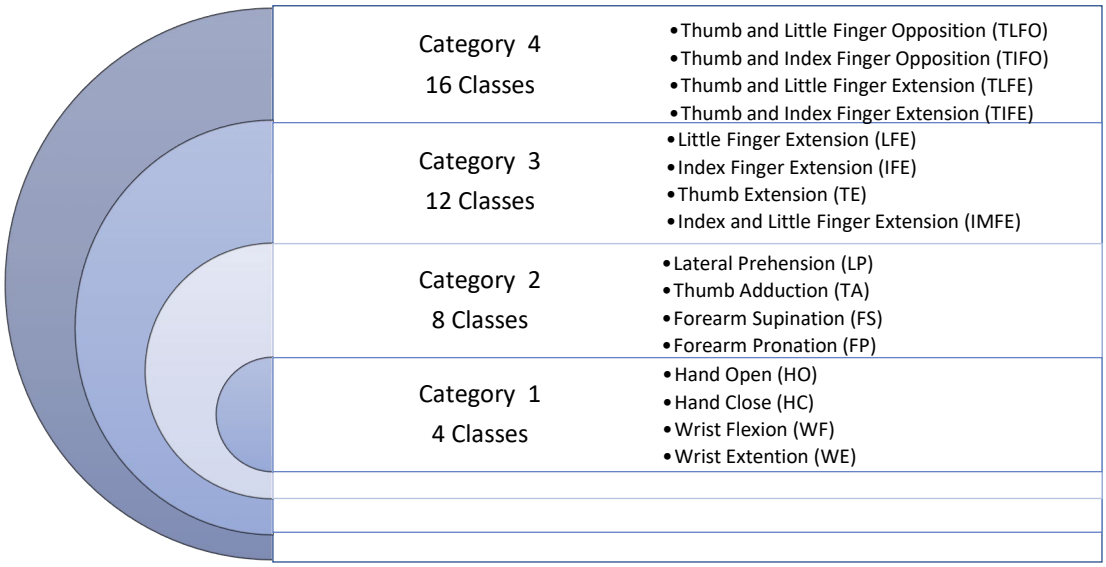


Figure 2. Movement Repertoire Expansion

III. RESULTS

This section evaluates the effects of movement repertoire size and electrode configuration on the performance of sEMG-based hand movement recognition. The experiments were conducted using four different movement repertoires comprising 4, 8, 12, and 16 movements, respectively, and three electrode configurations: wrist (12 channels), forearm (16 channels), and combined wrist and forearm (28 channels). The resulting accuracy, loss, precision, recall, and F1-score values are presented in Table 1.

For the 12-channel wrist electrode configuration, a consistent decrease was observed across all performance metrics as the movement repertoire increased. Test accuracy decreased from 96.18% for the 4-movement scenario to 81.30% when the movement repertoire was expanded to 16 movements. Similarly, the F1-score

decreased from 96.19% to 81.28%. These findings indicate that sEMG signals acquired exclusively from the wrist exhibit high discriminative capability for limited movement repertoires; however, classification performance decreases substantially as the movement repertoire expands.

Table 1. Classification performance of different electrode configurations across movement repertoire sizes, in terms of accuracy, loss, precision, recall, and F1-score.

Gesture Repertoire	Performance Metric	Electrode Combination		
		12	16	28
Category 1 (4 Classes)	Test Accuracy (%)	96.18	97.93	98.43
	Test Loss (%)	10.92	05.60	04.21
	Precision (%)	96.22	97.94	98.45
	Recall (%)	96.18	97.94	98.44
	F1 score (%)	96.19	97.93	98.44
Category 2 (8 Classes)	Test Accuracy (%)	88.31	87.48	95.35
	Test Loss (%)	33.72	31.18	13.56
	Precision(%)	88.35	87.63	95.40
	Recall (%)	88.29	87.48	95.37
	F1 score (%)	88.29	87.50	95.36
Category 3 (12 Classes)	Test Accuracy (%)	86.77	85.75	94.53
	Test Loss (%)	40.43	40.67	17.77
	Precision(%)	86.91	85.81	94.56
	Recall (%)	86.76	85.71	94.53
	F1 score (%)	86.76	85.70	94.53
Category 4 (16 Classes)	Test Accuracy (%)	81.30	81.31	92.31
	Test Loss (%)	57.01	54.20	24.78
	Precision(%)	81.48	81.41	92.36
	Recall (%)	81.28	81.27	92.30
	F1 score (%)	81.28	81.25	92.31

The 28-channel combined wrist–forearm electrode configuration achieved the highest classification performance across all movement repertoire scenarios. An accuracy of 98.43% was obtained for the 4-movement scenario, while an accuracy of 92.31% was maintained even in the most challenging 16-movement classification scenario. Furthermore, the performance advantage of the 28-channel configuration over the single-region configurations became increasingly pronounced as the movement repertoire expanded. While the difference in accuracy was approximately 2% in the 4-movement scenario, it increased to approximately 11% in the 16-movement scenario. These findings indicate that complementary sEMG data from different anatomical regions make a substantial contribution to the discrimination of movement classes, particularly when larger movement repertoires are considered.

Figure 3 compares the classification accuracies obtained using the three electrode configurations across different movement repertoire sizes. A consistent decrease in classification accuracy was observed across all electrode configurations as the movement repertoire size increased. Nevertheless, the 28-channel configuration combining wrist and forearm electrodes achieved the highest accuracy across all movement repertoires and maintained its

performance at a higher level than the other configurations. Notably, in the 16-movement scenario, the 28-channel configuration achieved an accuracy of 92.31%, whereas the wrist-only and forearm-only configurations achieved accuracies of 81.30% and 81.31%, respectively. Furthermore, the performance advantage of the 28-channel configuration over the single-region configurations became increasingly pronounced as the movement repertoire size increased. These findings indicate that complementary sEMG data from different muscle groups play an important role in maintaining classification performance as the movement repertoire expands.

Figure 4 presents the confusion matrices obtained for the 28-channel electrode configuration across movement repertoires comprising 4, 8, 12, and 16 movements. In the 4-movement scenario, the correct classifications are concentrated along the diagonal, while misclassifications remain negligible. As the movement repertoire size increases to 8 and 12 movements, a limited increase in inter-class confusion is observed; however, the diagonal structure remains largely preserved. In the most challenging 16-movement scenario, misclassifications become more frequent; nevertheless, these errors are concentrated between specific pairs of movements rather than being

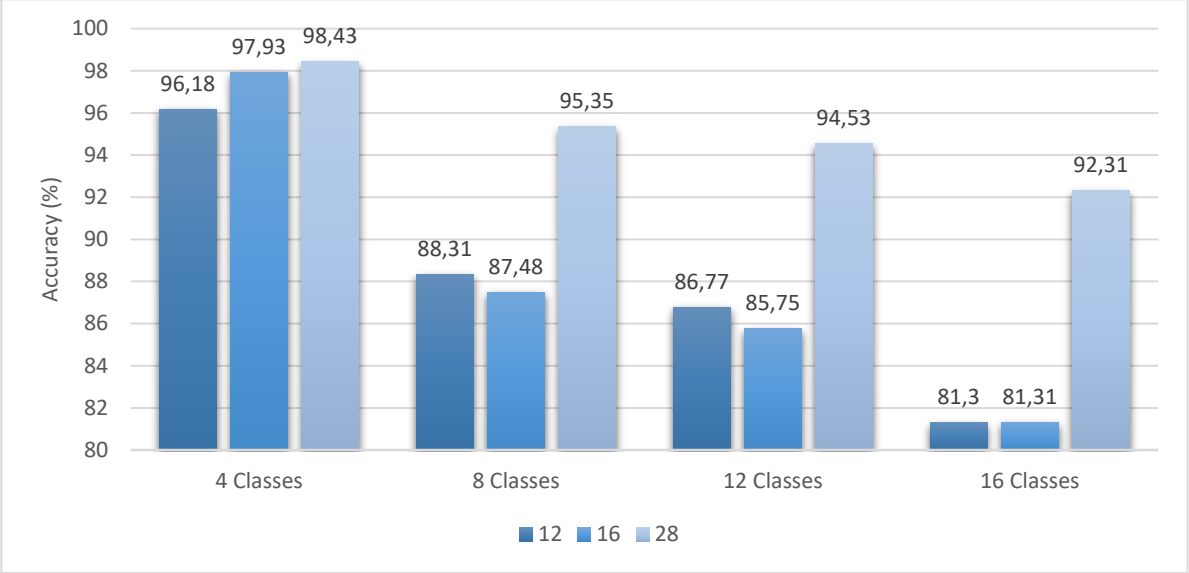


Figure 3. Comparison of classification accuracies obtained for different electrode configurations across different movement repertoire sizes.

randomly distributed across the classes. These findings indicate that the classification problem becomes increasingly complex as the movement repertoire size expands; despite this increased complexity, the 28-channel electrode configuration maintains a high level of discriminative capability even for larger movement repertoires.

This study systematically investigated the effects of movement repertoire size and electrode configuration on the classification performance of sEMG-based hand movement recognition systems. To this end, the GrabMyo dataset was used to evaluate four movement repertoires comprising 4, 8, 12, and 16 movements under three electrode configurations: wrist (12 channels), forearm (16 channels), and combined wrist and forearm (28 channels), all using the same CNN architecture. The results demonstrated that classification performance decreased across all electrode configurations as the movement repertoire size increased; however, the extent of this degradation varied by configuration.

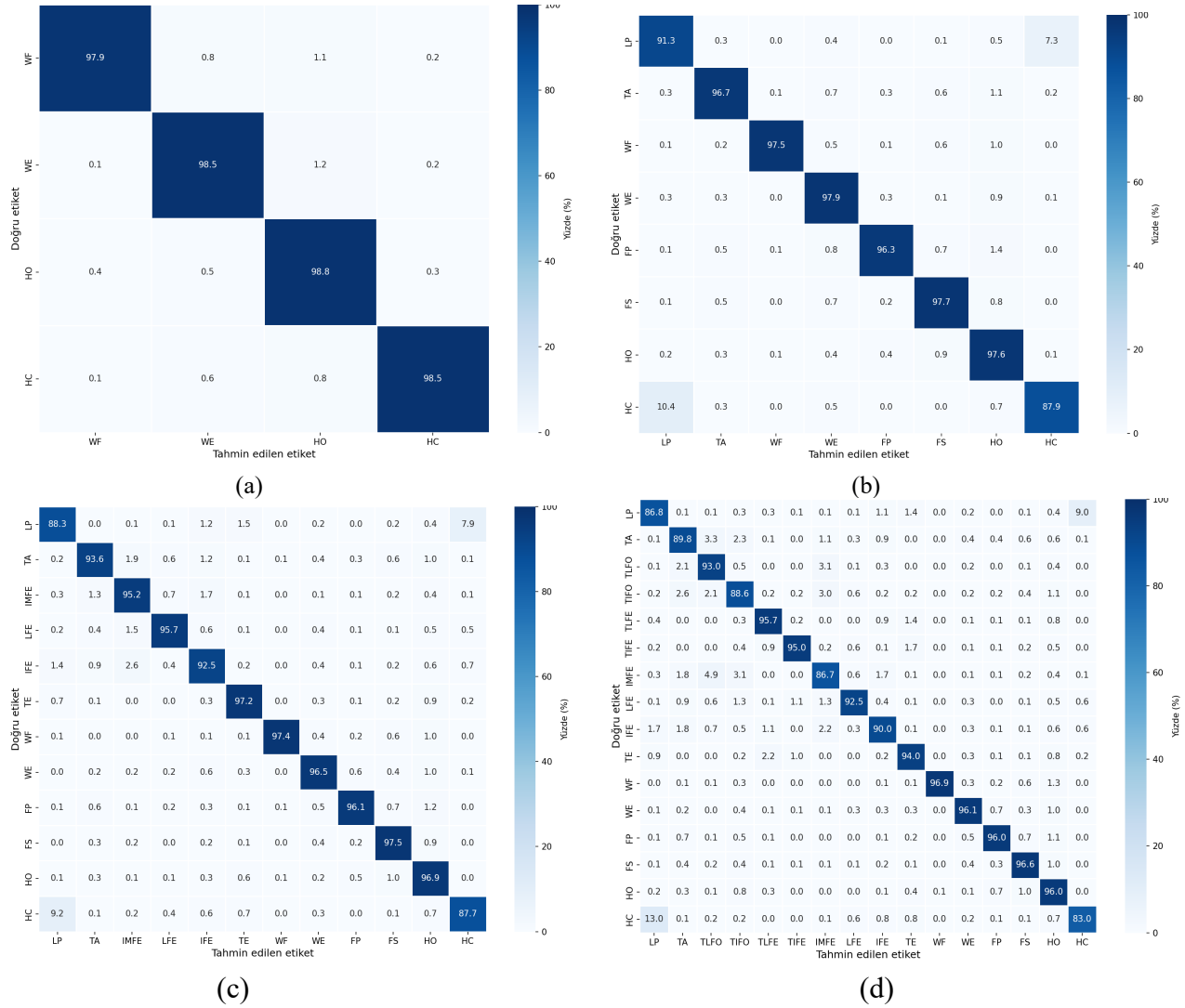


Figure 4. Confusion matrices obtained for the 28-channel electrode configuration across different movement repertoire sizes: (a) 4 movements, (b) 8 movements, (c) 12 movements, and (d) 16 movements.

III. CONCLUSION

The 28-channel combined wrist–forearm electrode configuration achieved the highest classification performance across all movement repertoires and, particularly for larger movement repertoires, maintained its performance more effectively than the other configurations. In contrast, the performance of the wrist-only and forearm-only configurations gradually converged as the movement repertoire increased. These findings indicate that complementary sEMG data from different anatomical regions make a substantial contribution to discriminating among movement classes, particularly when larger movement repertoires are considered. Overall, the findings highlight the importance of jointly considering movement repertoire size and electrode configuration in line with the requirements of different applications and provide useful insights for the design of scalable sEMG-based hand movement recognition systems.

REFERENCES

[1] W.N. Abdelrazik, et al., "Enhancing EMG-based gesture recognition for prosthetic control: Impact of dimensionality reduction on machine learning classifier," Sensors and Actuators A: Physical, 396: 117120, 2025.

[2] M.H. Zafar, S.K.R. Moosavi, and F. Sanfilippo, "Federated Learning-Enhanced Edge Deep Learning Model for EMG-Based Gesture Recognition in Real-Time Human–Robot Interaction," IEEE Sensors Journal, 25(5): 9139–9151, 2025.

[3] D. Leonardis, et al., "An EMG-Controlled Robotic Hand Exoskeleton for Bilateral Rehabilitation," IEEE Transactions on Haptics, 8(2): 140–151, 2015.

[4] C.L. Kok, et al., "Machine Learning-Based Feature Extraction and Classification of EMG Signals for Intuitive Prosthetic Control," Applied Sciences, 14(13): 5784, 2024.

[5] A.D. Mohapatra, A. Aggarwal, and R.K. Tripathy, "Automated Recognition of Hand Gestures From Multichannel EMG Sensor Data Using Time–Frequency Domain Deep Learning for IoT Applications," IEEE Sensors Letters, 8(6): 1–4, 2024.

[6] S.Tushir, N. Verma, and K. Kumar. "Exploring EMG Signals for Prosthetic Control and Gesture Recognition: Advancements, Applications, and Future Directions," 2024 International Conference on Computational Intelligence and Computing Applications (ICCICA), Samalkha, India, pp. 465-469, 2024.

[7] T. Triwiyanto, I.P.A. Pawana, and W. Caesarendra, "Deep learning approach to improve the recognition of hand gesture with multi force variation using electromyography signal from amputees," Medical Engineering & Physics, 125, 2024.

[8] M. A. Ozdemir, D. H. Kisa, O. Guren, A. Onan, and A. Akan, "EMG-based Hand Gesture Recognition using Deep Learning," 2020 Medical Technologies Congress (TIPTEKNO), Antalya, Turkey, pp. 1-4, 2020.

[9] X. Chen, et al., "Hand Gesture Recognition based on Surface Electromyography using Convolutional Neural Network with Transfer Learning Method," IEEE Journal of Biomedical and Health Informatics, 25(4): p. 1292–1304, 2021.

[10] A. Sultana, F. Ahmed, and M. Alam, "A systematic review on surface electromyography-based classification system for identifying hand and finger movements," Healthcare Analytics, 3: 100126, 2022.

[11] T. Fu, N. Jiang, C. He and J. He, "Wrist EMG-based Gestures Recognition for Finger and Wrist Motions," 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, pp. 1-4, 2024.

[12] P. Aarotale and A. Rattani, Machine Learning-based sEMG Signal Classification for Hand Gesture Recognition. .2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, pp. 6319-6326, 2024

[13] N. Jiang, A. Pradhan, & J. He, Gesture Recognition and Biometrics ElectroMyogram (GRABMyo) (version 1.1.0). PhysioNet, 2024RRID:SCR_007345, Available from: https://doi.org/10.13026/89dm-f662

[14] A. Pradhan, J. He, and N. Jiang, "Multi-day dataset of forearm and wrist electromyogram for hand gesture recognition and biometrics," Scientific Data, 9(1): 733, 2022.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Integrating NotebookLM into Flipped Mathematics Instruction: Insights from Pre-Service Teachers through the TPACK Framework

Aleksandar Milenković
Department of Mathematics and Informatics,
Faculty of Science, University of Kragujevac
Kragujevac, Serbia
aleksandar.milenkovic@pmf.kg.ac.rs
ORCID: 0000-0001-6699-8772

Nemanja Vučićević
Department of Mathematics and Informatics,
Faculty of Science, University of Kragujevac
Kragujevac, Serbia
nemanja.vucicevic@pmf.kg.ac.rs
ORCID: 0000-0002-4903-7280

Marina Svičević
Department of Mathematics and Informatics,
Faculty of Science, University of Kragujevac
Kragujevac, Serbia
marina.svicevic@pmf.kg.ac.rs
ORCID: 0000-0003-2791-3849

Abstract
Although many studies have examined the flipped classroom model and the use of AI chatbots in education, there is still little research on the educational potential of AI assistants such as NotebookLM in this context. This study investigates how pre-service mathematics teachers perceive the use of NotebookLM when planning mathematics lessons based on the flipped classroom model and preparing accompanying teaching materials. The study involved eleven fourth-year undergraduate mathematics students enrolled in a teacher education program. Using NotebookLM, they created presentations, video materials, and quizzes for mathematics lessons. After completing and presenting their lesson plans, the participants answered an open-ended questionnaire. Their responses were analyzed using thematic analysis and interpreted through the TPACK framework. The results show that the participants considered NotebookLM a useful tool for preparing instructional materials and for checking students' basic understanding before class. However, they also pointed out that teachers need to review, adapt, and improve the generated materials before using them in practice. The findings suggest that incorporating NotebookLM into pre-service mathematics teacher education can contribute not only to the development of digital competencies but also to strengthening critical AI literacy and to integrating the technological, pedagogical, and mathematical knowledge required for effective lesson planning.

Keywords: *NotebookLM, flipped classroom, pre-service mathematics teachers, TPACK*

I. INTRODUCTION

In recent years, innovation in mathematics education — and education more broadly — has been increasingly driven by the integration of technology, particularly artificial intelligence (AI). Unlike general-purpose chatbots such as Gemini, ChatGPT, and similar systems that generate responses based on knowledge acquired from vast amounts of internet data, NotebookLM generates responses grounded in the user's provided sources. (e.g., PDF files, text documents, web links, etc.). Consequently, NotebookLM generates responses grounded in the provided sources, substantially reducing the likelihood of AI "hallucinations." Moreover, generated responses include citations that allow users to trace information back to the relevant parts of the uploaded materials. A particularly valuable feature for educators is its ability to rapidly generate summaries, presentations, study guides, and even video or audio overviews (podcasts) from uploaded instructional materials. In addition, a growing body of educational research highlights the benefits of instructional approaches that enable students to construct knowledge as independently as possible. One such approach is the flipped classroom model. Although numerous studies have examined the implementation, advantages, and limitations of the flipped classroom, as well as the integration of AI chatbots into teaching and learning, research combining the flipped classroom model with AI-based technologies remains remarkably scarce in the contemporary educational literature.

The present study aims to address this evident gap in the literature by exploring pre-service mathematics teachers' perceptions of using NotebookLM for mathematics lesson planning and for developing instructional materials based on predefined teaching guidelines. To this end, fourth-year undergraduate mathematics students at the Faculty of Science, University of Kragujevac, designed mathematics lessons following the flipped classroom model with the support of the AI assistant NotebookLM. After implementing and presenting their lesson plans, the students completed an open-ended questionnaire. For this study, we analyzed their responses to two open-ended questions through the lens of the TPACK framework to investigate how, following this experience, pre-service mathematics teachers perceive and integrate technological, pedagogical, and mathematical content knowledge during the lesson-planning process. The contribution of this paper, therefore, lies in the use of a source-grounded assistant NotebookLM for mathematics lesson planning and in interpreting pre-service teachers' reflections through TPACK.

II. LITERATURE REVIEW

In this chapter, we briefly present the characteristics of the flipped classroom, the application of AI in pre-service teacher education, and the TPACK framework for integrating AI into education.

1. Flipped classroom and mathematics education

In their theoretical framework for the flipped classroom, Tomas, Evans, Doyle & Skamp [1] summarize the characteristics of the flipped classroom model compared with traditional instruction (Table 1). As can be observed, activities typically conducted in the classroom, such as learning through direct teacher exposition, are performed individually outside the classroom. In contrast, activities like practice exercises and research, which are typically carried out outside school, are conducted in the classroom under the teacher's guidance [2].

Table 1. Differences between traditional lectures and the flipped classroom (adapted from Tomas et al., 2019, p. 17)

Traditional Instruction	Flipped Classroom Model
The teacher provides a high level of guidance to students.	The student exercises a high level of autonomy.
Students acquire knowledge and skills through direct instruction delivered by the teacher in the classroom.	Students acquire knowledge through the provided materials before class.
During the lesson, an approach centered on content and activities addressing lower cognitive levels is adopted.	Emphasis is placed on the development of higher-order cognitive skills.
Acquired knowledge is applied after class through task analysis and practice exercises.	Through an active-learning approach in complex and novel contexts, knowledge is applied and expanded during class.

The results of the meta-analysis conducted by Sopamena [3] indicate that students who learned new mathematical concepts through the flipped classroom model achieved significantly better learning outcomes than those taught using traditional instructional approaches. One of the key advantages of the flipped classroom over traditional teaching is its ability to accommodate diverse students' learning needs and learning pace while optimizing instructional time [4]. The flipped classroom model overcomes the time constraints of conventional instruction by extending learning beyond the classroom, thereby giving students more time to engage in learning activities outside school. This approach also enables students to become more autonomous in managing their

own learning. Students acquire new concepts, rules, and procedures through instructional videos and other learning materials at their own pace. Rather than merely transmitting information, teachers facilitate learning by encouraging students to become active and independent participants in the learning process [4].

2. Generative AI and teacher education

In a review paper by Tan [5], which analyzed 22 empirical studies, the author identified seven outcomes regarding teachers' use of GenAI: improved learning, teacher knowledge, lesson design and implementation, higher-order thinking, AI literacy, self-worth and self-confidence, positive attitudes, and usage intentions. Nyaaba and Zhai [6] argue that teachers need to become familiar with generative AI tools during their initial studies as a core component of teacher education and continuous professional development programs, highlighting the importance of strengthening educators' skills in innovative assessment practices. Similarly, Panday-Shukla [7] notes that pre-service teachers hold both positive and skeptical views toward GenAI technologies, emphasizing that their general digital literacy is far more developed than their AI literacy. Consequently, pre-service teachers are not yet fully prepared to teach in AI-dominated learning environments, underscoring the need to design and implement additional professional development opportunities to learn to use GenAI tools. While AI tools can significantly boost creativity and confidence in lesson planning for future math teachers, special attention must be paid to the pedagogical approach to avoid superficial learning [8].

3. AI and TPACK

The Technological Pedagogical Content Knowledge (TPACK) framework, introduced by Mishra and Koehler [9], has become one of the most widely adopted theoretical models for understanding how teachers effectively integrate technology into instructional practice. Rather than viewing technological, pedagogical, and content knowledge as separate domains, the framework emphasizes their interaction and teachers' ability to combine these forms of knowledge to support meaningful learning experiences [10]. Within this perspective, technology is not regarded as an isolated instructional component but as a resource whose educational value depends on how it is aligned with pedagogical approaches and subject-specific content. Recent studies have increasingly employed the TPACK framework to examine teachers' readiness to adopt generative AI in educational settings. For example, Bautista et al. [11] reported that prospective teachers in the Philippines generally demonstrated positive attitudes toward integrating AI tools into their professional preparation, indicating a growing acceptance of AI-supported teaching practices. Similarly, Cun and Huang [12] investigated pre-service teachers' perceptions of generative AI through semi-structured interviews interpreted within the TPACK framework. Their findings revealed two dominant perspectives. On the one hand, participants viewed generative AI as a valuable resource that could support teachers in planning and delivering instruction. On the other hand, they expressed concerns regarding the responsible implementation of AI in the future.

4. Research question

Despite the growing body of research on generative AI in education, empirical evidence concerning NotebookLM remains scarce. Little is currently known about how pre-service mathematics teachers incorporate NotebookLM into lesson planning, especially within the flipped classroom model, or how they integrate its use with their technological, pedagogical, and mathematical content knowledge. The present study seeks to address this gap by examining pre-service mathematics teachers' experiences with NotebookLM through the conceptual lens of the TPACK framework.

R.Q. How do pre-service mathematics teachers perceive the use of AI assistant NotebookLM for planning mathematics lessons and creating instructional materials within the flipped classroom model?

III. METHOD

The study was conducted as part of a fourth-year undergraduate mathematics teacher education course involving pre-service mathematics teachers. As part of their instruction on the flipped classroom model, the students were introduced to the AI assistant NotebookLM and its capabilities for generating instructional videos, PowerPoint presentations, and quizzes. The flipped classroom instruction lasted for two weeks (8 classes), with two additional classes dedicated to explaining the features and usage of NotebookLM. Subsequently, they were assigned to develop instructional materials for planning and implementing mathematics lessons based on the flipped classroom model, and were encouraged to use NotebookLM. The resources used by the students to create video materials, PowerPoint presentations, and comprehension check questions included textbooks, problem books, and other instructional materials they typically use for lesson planning. The materials were then presented in class, discussed, and evaluated by both peers and the teacher. Direct interaction with primary or secondary students was not included in this course module. The study was conducted with 11 students enrolled in a course that introduces the flipped classroom model. None of the students had formal teaching experience as employed schoolteachers; however, as part of their pedagogical practice, they had completed an 8-week school placement in primary and secondary schools, where they observed and participated in classes under the guidance of a mathematics mentor teacher. Additionally, none of them had received formal training in integrating AI tools into teaching before being introduced to NotebookLM. The course syllabus was updated to incorporate this AI tool to generate instructional materials for the flipped classroom model; however, the use of NotebookLM was optional, allowing students to create their materials without AI or with alternative tools. Interestingly, all 11 students who participated in both the course and the study chose to use NotebookLM to develop their instructional materials. They created both learning materials for students to study independently at home before the mathematics lesson (Figures 1 and 2) and quizzes (Figure 3) to assess whether pupils had reviewed the assigned materials and acquired the fundamental concepts and knowledge related to the instructor's assigned topic.

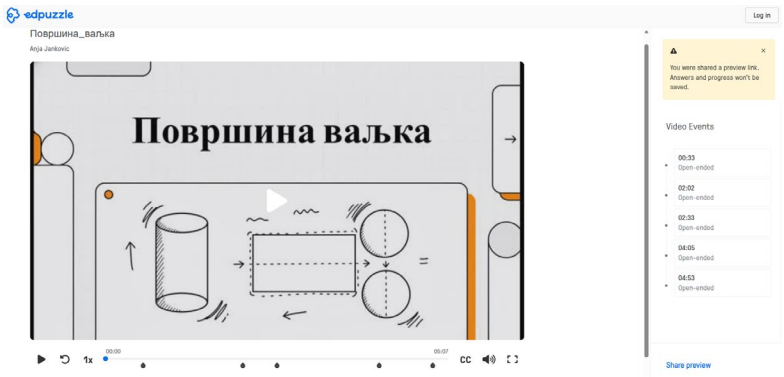


Figure 1. Video material for the lesson Surface Area of a Cylinder, created by a student in NotebookLM and then uploaded to the EdPuzzle platform for a flipped classroom lesson

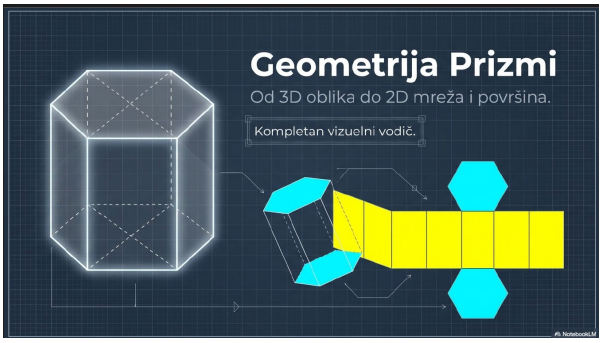


Figure 2. Presentation for the lesson Surface Area of a Prism, created by a student in NotebookLM for a flipped classroom lesson

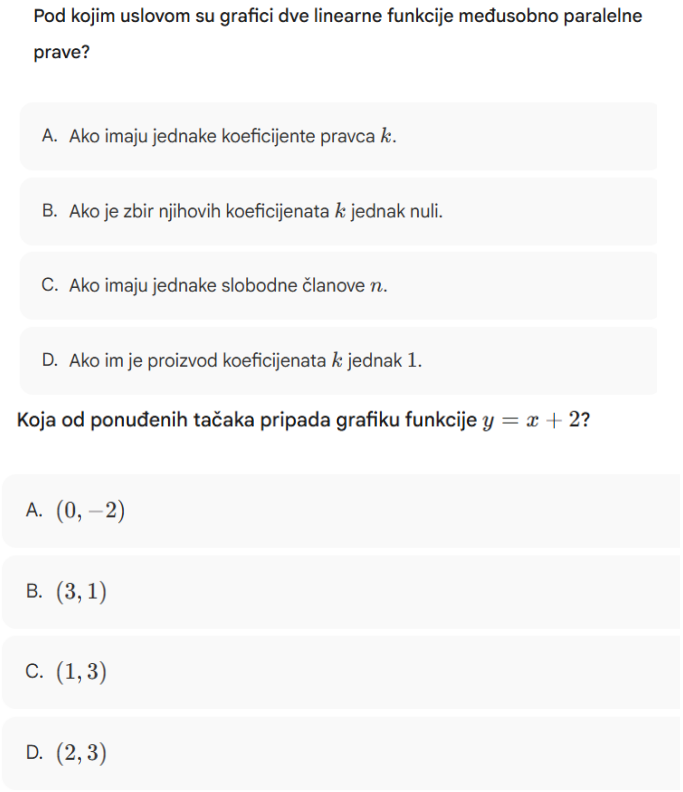


Figure 3. Quiz questions created by a student in NotebookLM to check content comprehension at home, before in-class learning for a flipped classroom model

The lesson topics corresponded to the final grade of elementary school (pupils aged approximately 15 years). The assigned topics covered geometry (similarity of triangles, Thales' theorem, polyhedra, solids of revolution, and their surface area and volume), algebra (linear equations and inequalities, and systems of linear equations), as well as an introductory topic in mathematical analysis (the linear function and its graph). Following the implementation of flipped mathematics lesson plans using the AI tool NotebookLM, pre-service teachers completed an open-ended questionnaire to provide insights into their perceptions of using this technology in

lesson planning and in the methodical design of instructional materials. For this study, we analyzed the responses to the following questions:

How do you evaluate the pedagogical quality of the knowledge assessment tests generated by NotebookLM?

How did using NotebookLM affect your perception of the teacher's role in a flipped classroom? Do you see AI as a collaborator that frees up time for more creative work with students, or as a factor that could uniformize and impoverish the instructional process?

Participation in the study was voluntary. Students were informed about the purpose of the research, and all responses were anonymized before analysis. The study was conducted after the end of the semester, so the students were under no obligation to answer the questions, and their participation had no bearing on their final course grades. The responses were analyzed using thematic analysis, following the guidelines outlined by Braun and Clarke [13]. First, the researchers independently familiarized themselves with the complete dataset, and initial codes were generated inductively from the participants' responses without assigning them to predefined TPACK categories. The researchers then compared related codes, grouped conceptually similar codes into broader categories, and iteratively refined the emerging themes. TPACK was not used as a predefined coding scheme during the initial coding stage. Instead, after the themes were established from the data, they were interpreted using the TPACK framework to examine the types of knowledge reflected in participants' responses. For this study, Technological Knowledge (TK) refers to familiarity with NotebookLM's capabilities, functionalities, practical applications, and potential limitations. Pedagogical Knowledge (PK) encompasses an understanding of learning processes, instruction, assessment, student engagement, and the teacher's role, independent of specific technology or mathematical content. Content Knowledge (CK) refers to the understanding of mathematical concepts, processes, and core mathematical activities that should be embedded in the developed materials, including conceptual understanding, reasoning, problem-solving, and real-world applications. Technological Pedagogical Knowledge (TPK) was identified when participants linked specific features or limitations of NotebookLM to concrete pedagogical considerations. Technological Content Knowledge (TCK) was recognized when participants connected the use or constraints of NotebookLM to how mathematical content is represented and processed. Pedagogical Content Knowledge (PCK) was coded when participants addressed pedagogically sound approaches to teaching mathematics in which the technological component played a secondary role. Finally, Technological Pedagogical Content Knowledge (TPACK) was identified when participants simultaneously aligned the technical capabilities or constraints of NotebookLM with pedagogical goals and the unique demands of mathematics learning. This operationalization was used to distinguish between statements that merely expressed a positive or negative stance toward artificial intelligence and those reflecting integrated professional knowledge.

IV. RESULTS AND DISCUSSION

The thematic analysis of student responses revealed that students generally perceive NotebookLM as a useful technological tool for preparing instructional materials within the flipped classroom model. At the same time, their responses indicate a strong critical awareness regarding the pedagogical limitations of this tool. Rather than unquestioningly accepting AI-generated content, students emphasize the importance of teachers' professional and critical judgment, highlighting the need to adapt, refine, and contextualize NotebookLM-generated materials. The identified themes were interpreted through the TPACK framework, providing insight into how pre-service teachers integrate technological, pedagogical, and mathematical content knowledge during lesson planning.

Theme 1: NotebookLM Effectively Supports Basic Knowledge Verification but Has Limited Capacity for Assessing Conceptual Understanding. The first and most prominent theme relates to the assessment of the pedagogical quality of tests generated by NotebookLM. Nearly all pre-service teachers (9 of them) believe that the tool successfully generates questions that verify whether students have reviewed the instructional materials and acquired the foundational concepts necessary for in-class activities. They describe the generated questions as "solid (P1)," "well-aligned with the presentation (P4)," and "suitable for basic knowledge checking (P5)." Seven students emphasize that this specific type of verification is an important function of the flipped classroom model—providing the teacher with information about whether students are prepared for subsequent classroom activities. In doing so, the students demonstrated that they evaluated the questions through an appropriate lens, analyzing them specifically within the context of their implementation in a flipped classroom model. One student noted that the generated questions are "solid for basic verification and recall, which is essentially the goal before students enter the classroom (P1)." In contrast, another student stated that the quiz "serves as an excellent check of whether students viewed the presentation and acquired what was necessary for that particular lesson (P7)." Similar views emerged across the majority of respondents, indicating a high degree of consensus regarding NotebookLM's core function during the student pre-class preparation phase. Conversely, 8 students simultaneously draw a clear distinction between verifying rote knowledge and evaluating true mathematical understanding. In their view, the questions generated by the tool primarily require recalling definitions, recognizing formulas, or remembering information explicitly presented in the slides. One respondent stated that students could successfully pass the test "simply by memorizing details they had previously seen (P8). It is particularly notable that students spontaneously differentiated between reproductive knowledge and conceptual understanding. They highlighted that high-quality assessment in mathematics instruction must incorporate tasks requiring analysis, knowledge integration, logical reasoning, and the application of mathematical concepts to novel situations. Such responses demonstrate a developed pedagogical awareness of the nature of mathematical comprehension, indicating that students do not equate successful quiz completion with genuine learning. Viewed through the TPACK framework, this theme primarily reflects the Technological Pedagogical Knowledge (TPK) component. Pre-service teachers demonstrate an understanding that a specific technology can effectively support a particular pedagogical goal—checking student readiness before class—while simultaneously recognizing the technology's limits when evaluating deeper mathematical comprehension.

Theme 2: The Pedagogical Value of AI Materials Emerges Only After Teacher Intervention

The second dominant theme concerns the need for teachers to adapt and refine the questions generated by NotebookLM. The majority of pre-service teachers (7 of them) believe that the generated quizzes serve as a solid starting point, but are not ready for immediate classroom application without further modification. Students reported rephrasing certain items, adding simpler problem-solving tasks, or incorporating questions that require applying knowledge in novel situations. One student noted that she "had to manually modify several questions and add problem-based tasks that require applying knowledge in a new context (P2)." At the same time, another mentioned reshaping the questions to demand analysis and application rather than mere recall. Interestingly, students do not view this additional refinement as a shortcoming of NotebookLM, but rather as an integral component of a teacher's professional practice. In other words, they perceive AI as a tool that accelerates the preparation of instructional materials, while the ultimate responsibility for quality rests with the mathematics teacher.

Within the TPACK framework, this theme represents the clearest illustration of Technological Pedagogical Content Knowledge (TPACK) in action. Pre-service teachers do not merely evaluate the tool's technological capabilities; rather, they synthesize AI affordances with pedagogical principles and the specific demands of mathematical content. Their pedagogical interventions demonstrate an understanding that high-quality

assessment in mathematics requires carefully designed tasks that foster reasoning, problem-solving, and conceptual understanding—elements that highlight why technology cannot replace teacher expertise.

Theme 3: NotebookLM Transforms the Role of the Teacher, but Does Not Replace It

Responses to the second questionnaire question provided rich insights into how pre-service teachers perceive the transformation of their professional role in an environment that integrates generative AI. Regardless of varying degrees of enthusiasm for NotebookLM, nearly all students (8) maintain that AI cannot replace the mathematics teacher. Six students described NotebookLM as a collaborator that handles routine technical tasks, allowing the teacher to devote more time to lesson planning, direct interaction with students, and the design of creative learning activities. They particularly emphasize the time savings achieved in creating presentations, video materials, and quizzes. One student noted that NotebookLM "freed up a vast amount of time that could be redirected toward creatively designing in-class activities (P3)." At the same time, another highlighted that the tool allows her to focus on "the essence of the profession—more creative engagement with students and tailoring instruction to their individual needs (P11). Simultaneously, pre-service mathematics teachers stress that the responsibility for quality control of instructional materials must remain with the teacher. Five respondents indicated that AI-generated materials should be carefully vetted before implementation, while some cautioned that uncritical reliance on NotebookLM could lead to a decline in instructional quality. One student observed that NotebookLM "merely helps prepare materials faster, but the teacher must judge whether it is of sufficient quality (P6)." At the same time, another emphasized that "the teacher's role remains pivotal because AI cannot function independently without constant oversight and correction (P9)." These findings suggest that pre-service teachers do not view AI as a replacement for the educator, but rather as an instrument that alters the nature of teaching practice. Instead of spending the majority of their time on technical material preparation, the teacher becomes an instructional designer, mentor, and professional decision-maker in lesson planning. Viewed through the TPACK framework, this theme illustrates that pre-service teachers are developing an understanding that successful technology integration requires more than digital proficiency alone; it demands the ability to thoughtfully embed tools into the instructional process while upholding professional pedagogical responsibility.

Theme 4: Critical AI Literacy as a Prerequisite for the Effective Integration of NotebookLM

Although pre-service teachers generally hold a positive attitude toward NotebookLM, a significant number of responses indicate a clear awareness of the potential risks associated with its uncritical application. Interestingly, however, students almost unanimously maintain that the problem lies not in the technology itself but in how the teacher uses it. Four student respondents noted that instructional standardization or homogenization could occur only if teachers uncritically accepted everything generated by NotebookLM. One student noted that AI could impoverish the educational process only if a teacher "blindly copies whatever the AI outputs (P10)," while other respondents cautioned that automatically generated materials must not become the sole source of instructional activities. These responses demonstrate that pre-service teachers do not view technology as an end in itself, but rather as a means whose value hinges directly on teacher competence. Their answers point to the development of critical AI literacy—specifically, the teacher's capacity to evaluate the quality of AI-generated content, recognize its limitations, and adapt it to the needs of a specific instructional setting. From the perspective of the TPACK framework, this theme confirms that the successful integration of AI technologies requires the simultaneous development of technological, pedagogical, and content (specifically mathematical) competencies. Only through their mutual synthesis will teachers be equipped to make sound didactic decisions when planning mathematics instruction using AI technology.

General Interpretation of Results Through the TPACK Framework

Taken as a whole, the results suggest that pre-service mathematics teachers do not perceive NotebookLM as an autonomous solution for instructional planning, but rather as a technological resource whose educational value hinges directly on the teacher's pedagogical and domain-specific knowledge. Their responses consistently emphasize the necessity for critical evaluation, adaptation, and refinement of AI-generated instructional materials. It is particularly noteworthy that students do not merely debate whether NotebookLM is useful; instead, they reflect deeply on the specific contexts, instructional goals, and pedagogical strategies through which it should be employed. Such reflection extends beyond the mere adoption of a new technology. It serves as a clear indicator of the development of integrated professional knowledge, as described by the TPACK framework, the harmonious synthesis of technological, pedagogical, and content competencies. Consequently, the findings suggest that embedding NotebookLM into lesson-planning activities within initial teacher education can not only enhance pre-service teachers' digital skills but also cultivate their capacity to critically and thoughtfully integrate generative artificial intelligence into mathematics instruction. The results of our research align with the pedagogical principles of the flipped classroom. Namely, pupils emphasize that students should study the material thoroughly at home and master foundational concepts before class [2],[4]. Accordingly, our students perceived NotebookLM-generated quizzes as appropriate tools for verifying pupils' readiness rather than for assessing higher-order mathematical thinking. They noted that the generated materials were visually appealing and engaging. These findings are also consistent with previous studies reporting that pre-service teachers generally perceive generative AI as a valuable assistant rather than a replacement for teacher expertise [5],[7],[8]. Participants in this study emphasized that instructional quality depends on teachers' judgment and their ability to adapt AI-generated materials to specific mathematical content and classroom activities. From the perspective of the TPACK framework, the results suggest that participants did not perceive technology as an isolated competence. They viewed NotebookLM as a technological resource whose educational value depends on pedagogical decisions and content knowledge. This supports the central assumptions of the TPACK framework [9] and suggests that learning to work with an AI assistant should focus not only on developing digital skills but also on cultivating integrated professional knowledge.

V. CONCLUSION

This study explored pre-service mathematics teachers' perceptions of using NotebookLM for planning mathematics lessons within the flipped classroom model through the lens of the TPACK framework. The findings suggest that the participants perceived NotebookLM as a useful tool for developing instructional materials, including presentations, video content, and quizzes, particularly for supporting pupils' pre-class preparation. At the same time, they emphasized that high-quality mathematics instruction cannot rely solely on AI-generated materials. Instead, effective lesson planning requires teachers to critically evaluate, adapt, and enrich the generated content by drawing on their pedagogical expertise and mathematical knowledge. The findings also underline the importance of developing critical AI literacy alongside digital competencies in initial teacher education. Providing pre-service teachers with opportunities to work with source-based generative AI tools, such as NotebookLM, may not only improve their technological skills but also foster more reflective, pedagogically informed, and critically aware approaches to integrating artificial intelligence into mathematics education. This study has several limitations. First, the sample consisted of only eleven pre-service mathematics teachers from a single teacher education program, limiting the transferability of the findings. Second, the results were based exclusively on participants' self-reported perceptions. Finally, the study examined only short-term experiences with NotebookLM, which may have been influenced by the researcher, subject to social desirability bias, and limited to a narrow two-question dataset. Future research should therefore investigate the long-term integration of source-based generative AI tools in teacher education and compare NotebookLM with other generative AI systems used for instructional planning.

ACKNOWLEDGEMENT

Authors are supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, under Contract No. 451-03-34/2026-03/ 200122.

REFERENCES

[1]. L. Tomas, N. S. Evans, T. Doyle, and K. Skamp, "Are first year students ready for a flipped classroom? A case for a flipped learning continuum," International Journal of Educational Technology in Higher Education, vol. 16, no. 1, pp. 1–22, 2019.

[2]. S. C. Kong, "Developing information literacy and critical thinking skills through domain knowledge learning in digital classrooms: An experience of practicing flipped classroom strategy," Computers & Education, vol. 78, pp. 160–173, 2014, doi: 10.1016/j.compedu.2014.05.009.

[3]. P. Sopamena, K. Sangadji, D. Riaddin, S. Kaliky, and G. Assagaf, "Effectiveness of flipped classroom model on mathematics achievement at the university level: A meta-analysis study," International Journal of Instruction, vol. 16, no. 1, pp. 767–780, 2023, doi: 10.29333/iji. 2023.16143a.

[4]. J. Bergmann and A. Sams, "Flip Your Classroom: Reach Every Student in Every Class Every Day. " Washington, DC, USA: International Society for Technology in Education, 2012.

[5]. Q. Tan, "Reimagining teacher development in the era of generative AI: A scoping review," Teaching and Teacher Education, vol. 168, Art. no. 105236, 2025.

[6]. M. Nyaaba and X. Zhai, "Generative AI professional development needs for teacher educators," Journal of AI, vol. 8, no. 1, pp. 1–13, 2024, doi: 10.61969/jai. 1385915.

[7]. P. Panday-Shukla, "Exploring generative artificial intelligence in teacher education," Teaching and Teacher Education, vol. 165, Art. no. 105088, 2025.

[8]. S. Oh, "Integration of MATH41 and generative AI in pre-service mathematics teacher education: An empirical study on lesson design competency," IEEE Access, 2025.

[9]. P. Mishra and M. J. Koehler, "Technological pedagogical content knowledge: A framework for teacher knowledge," Teachers College Record, vol. 108, no. 6, pp. 1017–1054, 2006.

[10]. M. Koehler and P. Mishra, "What is technological pedagogical content knowledge (TPACK)?" Contemporary Issues in Technology and Teacher Education, vol. 9, no. 1, pp. 60–70, 2009.

[11]. A. Bautista, C. Estrada, A. M. Jaravata, L. M. Mangaser, F. Narag, R. Soquila, and R. J. Asuncion, "Preservice teachers' readiness towards integrating AI-based tools in education: A TPACK approach," Educational Process: International Journal, vol. 13, no. 3, pp. 40–68, 2024.

[12]. A. Cun and T. I. N. G. Huang, "Generative AI and TPACK in teacher education: Pre-service teachers' perspectives," in "Exploring New Horizons: Generative Artificial Intelligence and Teacher Education", pp. 62–75, 2024.

[13]. V. Braun and V. Clarke, "Using thematic analysis in psychology," Qualitative Research in Psychology, vol. 3, no. 2, pp. 77–101, 2006.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

SpecGuard: Multi-Layer Quality Assessment of Natural-Language FPGA Requirements

Anton Striapunin
Department of Cybersecurity and Intelligent Information Technologies, National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
a.o.striapunin@csn.khai.edu

Abstract
Natural-language specifications remain an important input artifact in the development and verification of FPGA-based components, including safety-critical airborne electronic hardware. Although natural language processing for requirements engineering has been widely studied for software requirements, open FPGA specifications remain underexplored as a target for automated quality assessment. This paper presents SpecGuard, a multi-layer framework for assessing the quality of natural-language FPGA requirements before downstream verification and formalization activities. SpecGuard combines rule-based requirements-smell detection, linguistic metrics, composite quality scores, and PASS/WARN/FAIL gating rules. The smell catalog is mapped to ISO/IEC/IEEE 29148 quality characteristics and prioritized with respect to DO-254-oriented verification concerns. A pilot study of 64 requirements from the open-source CVA6 RISC-V specification evaluates the distribution of smells, readability, and syntactic complexity metrics, the behavior of scores, and sensitivity to seeded defects. The results show that SpecGuard can process a mature industrial open-source FPGA specification and produce a diagnostic baseline for future comparative studies. The study is limited by the single-corpus design, the author-proposed severity weights, which are not yet empirically calibrated, and the absence of independent expert annotation.

Keywords: requirements engineering, requirements smells, FPGA, DO-254, NLP4RE, quality assessment, verification.

I. INTRODUCTION

Natural-language requirements remain the primary means of expressing the functional and non-functional properties of engineering systems. In safety-critical domains, the quality of these requirements directly affects the feasibility of subsequent review, verification, traceability, and certification-related activities. ISO/IEC/IEEE 29148:2018 defines several characteristics of a well-formed requirement, including unambiguity, completeness, singularity, feasibility, verifiability, correctness, and conformance to the prescribed style [1]. However, the standard defines desired properties rather than operational procedures for automated measurement. For airborne systems, the situation is further complicated by the separation between software assurance and airborne electronic hardware assurance. Software components are typically considered in relation to RTCA DO-178C, while FPGA, ASIC, and PLD components fall under the design-assurance guidance of RTCA DO-254 [2], [3]. In this context, natural-language specifications of FPGA components are early-life-cycle inputs that influence subsequent verification activities, including test development, assertion synthesis, traceability analysis, and the review of design-assurance evidence. By FPGA requirements, we mean natural-language requirements for HDL-implemented digital hardware (soft IP cores and FPGA/ASIC designs described in VHDL or Verilog), as distinct from requirements for software running on a fixed microprocessor. The distinction is regulatory rather than stylistic: such hardware falls under DO-254 design-assurance guidance, whereas microprocessor-hosted software falls under DO-178C, even though both share the natural-language specification medium. In this paper,

the unit of analysis is the HDL design at the core or IP level, not a full board- or system-level integration; system-level hardware/software interaction is out of scope and noted as future work. Natural language processing for requirements engineering has become an active research field. Existing work covers tasks such as requirements classification, ambiguity detection, traceability, and quality assessment [4], [5]. At the same time, most existing work focuses on software requirements or specific regulated domains, while natural-language FPGA specifications remain underrepresented as a research object. This creates a methodological and empirical gap: it is unclear which quality indicators are suitable for FPGA specifications and how they behave in realistic industrial open-source artifacts. This paper introduces SpecGuard, a lightweight multi-layer framework for automated assessment of natural-language FPGA requirements. The framework is not intended to prove certification readiness or replace expert review. Instead, it acts as a diagnostic quality gate, identifying rule-based quality issues before requirements are used in downstream verification or formalization activities. The main contributions are: (1) a multi-layer framework for assessing natural-language FPGA requirements; (2) a DO-254-oriented adaptation of requirements-smell analysis; and (3) a pilot empirical evaluation on the CVA6 RISC-V specification, including smell distribution, linguistic metrics, composite scores, gate decisions, and a seeded-fault sensitivity check.

II. RELATED WORK

1. Requirements Quality and Quality Characteristics

The normative basis for requirements quality is commonly associated with ISO/IEC/IEEE 29148:2018, which defines quality characteristics of individual requirements and requirements specifications [1]. These characteristics provide a useful vocabulary for discussing the quality of requirements, but they do not directly define measurable indicators or automated procedures. Systematic studies of requirements quality indicate that the field remains methodologically fragmented. Different studies use different attribute definitions, datasets, measurement procedures, and validation strategies [6]. This motivates the use of a transparent and diagnostic framework rather than a single scalar quality score.

2. Requirements Smells

Requirements smells provide an operational bridge between abstract quality characteristics and detectable linguistic symptoms. Femmer et al. framed smells as concrete indicators of potential quality problems in natural-language requirements [7]. Later studies demonstrated that smell catalogs can be adapted to specific domains and industrial contexts, including finance, safety-critical defense systems, and aircraft systems [8–10]. For the present work, requirements smells are attractive because they are interpretable, deterministic, and actionable. Each detected smell can be associated with a specific text fragment, a quality characteristic, and a recommendation for review.

3. Quantitative Quality Metrics

Several approaches aggregate quality indicators into composite metrics. Some frameworks use rule-based compliance with requirement-writing guidelines, while others focus on testability, readability, or syntactic complexity [11], [12]. However, a single aggregate score may hide diagnostic information. SpecGuard, therefore, combines composite scores with explicit smell reports and gating decisions. Linguistic metrics such as Flesch Reading Ease, Flesch-Kincaid Grade Level, and Mean Dependency Length provide an additional layer of analysis [13], [14]. They do not directly measure correctness or verification readiness, but they can reveal readability and syntactic-complexity patterns that smell detectors do not capture.

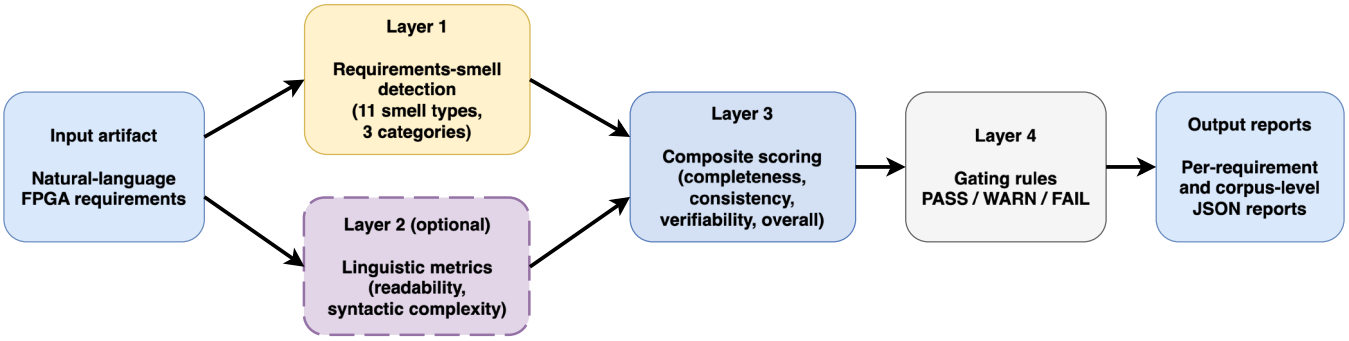
4. FPGA Specifications and Regulated Environments

Existing work on regulated requirements engineering has mainly focused on software-intensive systems and standards such as DO-178C, IEC 62279, and IEC 62304 [15], [16]. FPGA-specific natural-language specifications are less visible in NLP4RE research, despite the importance of FPGA-based components in critical systems. Open-source hardware projects such as CVA6 offer a practical way to study industrial-grade FPGA documentation without relying on proprietary artifacts. CVA6 has been used as a benchmark in hardware verification research, including studies on HDL and assertion synthesis [17]. However, its natural-language requirements have not been systematically analyzed as part of a requirements-quality corpus.

III. SPECGUARD FRAMEWORK

1. Framework Overview

SpecGuard is organized as a four-layer pipeline. Layer 1 detects requirements smells in individual requirements. Layer 2 computes linguistic metrics, including readability and syntactic-complexity indicators. Layer 3 aggregates smell and metric information into composite quality indicators. Layer 4 assigns a PASS, WARN, or FAIL decision based on the gating rules. The purpose of this layered architecture is to avoid two extremes. Isolated low-level metrics do not provide an actionable per-requirement verdict (PASS / WARN / FAIL) that an engineer can act on during review. Conversely, a single aggregate score hides the nature of the detected problem. SpecGuard, therefore, preserves diagnostic information while still producing a practical gate decision. Figure 1 summarises the pipeline overview.



Layer 2 (dashed border) is an optional component that requires textstat and spaCy.
When enabled, its linguistic metrics are aggregated into Layer 3 alongside smell signals from Layer 1.

Figure 1. Pipeline overview of the SpecGuard framework. Layer 2 (dashed border) is an optional component; when enabled, its linguistic metrics are aggregated with smell signals from Layer 1 into Layer 3.

2. Smell Catalog

The current SpecGuard catalog contains 11 smell types grouped into three categories. Table 1 records a diagnostic ISO/IEC/IEEE 29148 association, not a one-to-one scoring map. The implementation penalizes verifiability for ambiguity, vagueness, subjectivity, comparative, and missing_unit, and completeness for placeholder, implicit_reference, and optionality. Weakness, non-verifiable, and negative statements remain report-only diagnostics; consistency is determined using separate text heuristics. Severity, therefore, affects the

score only for smells in the implemented scoring set. The mappings and severity values are author-assigned and require expert calibration (Section VI).

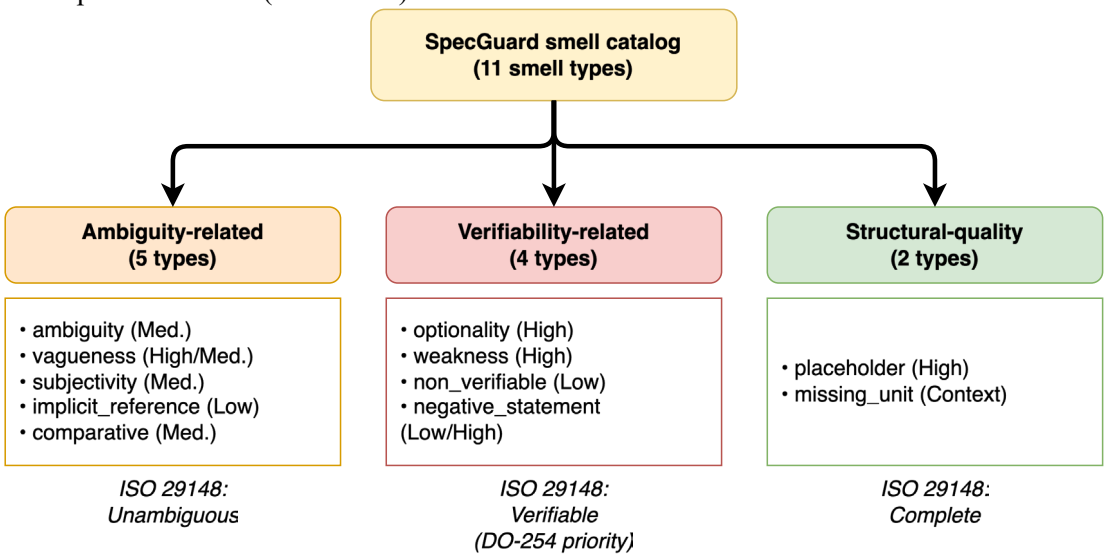


Figure 2. Hierarchy of the SpecGuard smell catalog: three categories aligned with ISO/IEC/IEEE 29148 quality characteristics.

Table 1. REQUIREMENTS-SMELL CATALOG USED IN SPECGUARD Cat. is the catalog grouping (ambiguity/verifiability / structural) used to organize detectors; Quality is the ISO/IEC/IEEE 29148 quality characteristic that the smell violates.

Type	Cat.	Quality	Sev.
ambiguity	Amb.	Unambiguous	Med.
vagueness	Amb.	Unambiguous	High/Med.
subjectivity	Amb.	Unambiguous	Med.
implicit_reference	Amb.	Unambiguous	Low
comparative	Amb.	Unambiguous	Med.
optionality	Verif.	Verifiable	High
weakness	Verif.	Verifiable	High
non_verifiable	Verif.	Verifiable	Low
negative_statement	Verif.	Verifiable	Low/High
placeholder	Struct.	Complete	High
missing_unit	Struct.	Complete/Verif.	Context

3. Linguistic Metrics and Gating Rules

The linguistic layer computes Flesch Reading Ease, Flesch-Kincaid Grade Level, Mean Dependency Length, token count, sentence count, average sentence length, and lexical density. These metrics are not interpreted as

direct evidence of correctness. Instead, they serve as supplementary indicators of cognitive load and syntactic complexity.

SpecGuard computes completeness, consistency, verifiability, and overall = 0.30 × completeness + 0.25 × consistency + 0.45 × verifiability. Scores are bounded to [0, 1] and rounded to three decimals; absence of a smell does not necessarily yield 1.0. Completeness starts at 0.70, adds 0.20 for modal commitment and 0.10 for an uppercase subject identifier, then subtracts mapped penalties; placeholders incur the normal penalty plus an additional 1.5 × severity. Verifiability starts at 0.50, adds 0.30 for a measurable criterion, and up to 0.20 for modal strength, then subtracts mapped penalties. Consistency starts at 1.0 and subtracts 0.20 for co-occurring positive and negative mandatory forms and 0.15 for mixed mandatory and optional modality. Cross-requirement consistency is outside the present gate. Layer 4 assigns PASS for overall ≥ 0.75, WARN for 0.50 ≤ overall < 0.75, and FAIL otherwise. PASS means only that the weighted score meets the threshold; smells may remain, and the requirement is not thereby correct or certification-ready. The weights are author-proposed, following prior composite-quality practice [18], and require expert calibration and sensitivity analysis.

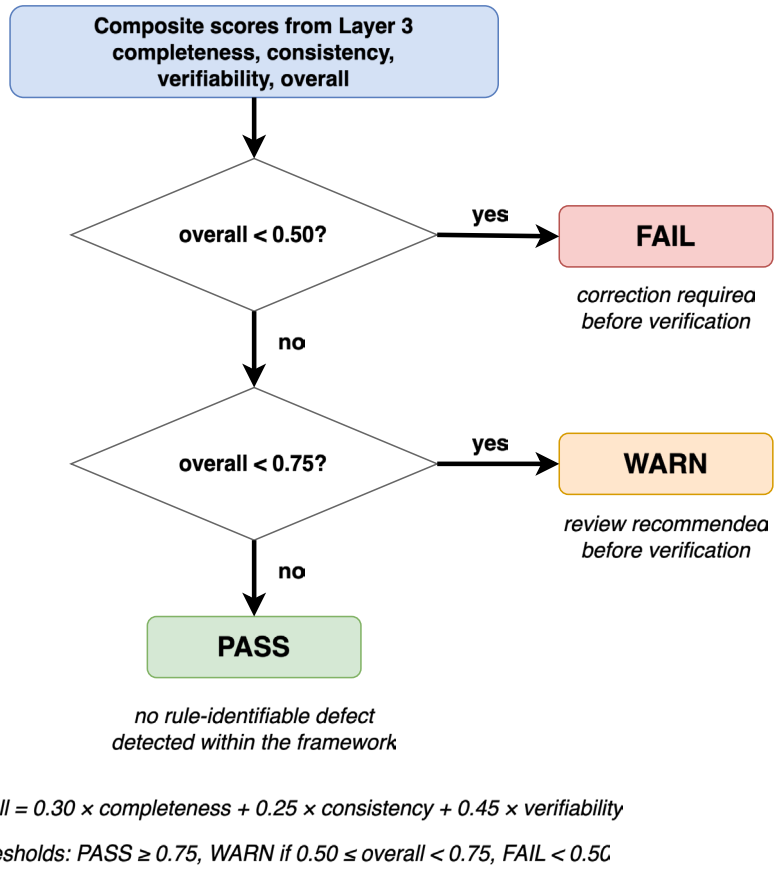


Figure 3. Layer 4 gating decision flow. Thresholds applied to the composite overall score: PASS ≥ 0.75, WARN if 0.50 ≤ overall < 0.75, FAIL < 0.50.

IV. IMPLEMENTATION AND STUDY DESIGN

1. Prototype Implementation

SpecGuard was implemented as a Python package with a deterministic core and optional extensions. The core implements smell detection, composite scoring, and gating decisions using standard Python facilities and curated lexical patterns. The optional linguistic layer uses Textstat for readability metrics and spaCy for dependency parsing [19], [20]. The tool produces per-requirement and corpus-level JSON reports to support reproducibility. SpecGuard uses no machine-learning models. Smell detection is deterministic, based on regular expressions and curated lexicons derived from Femmer et al. [7]; the optional linguistic layer uses spaCy solely for dependency parsing and textstat for classical readability formulas. This is a deliberate design choice favoring transparency, auditability, and amenability to future DO-330 tool-qualification analysis, rather than the higher recall an opaque machine-learning classifier might offer. The current prototype implements 11 active smell detectors and includes automated tests for core detector behavior, scoring, pipeline integration, and selected auxiliary modules.

2. Research Questions and Corpus

The pilot study is organized around four research questions: RQ1 asks how detected requirements smells are distributed across the CVA6 specification; RQ2 examines the linguistic-metric profile of the CVA6 requirements; RQ3 investigates whether the layers of SpecGuard provide complementary signals; and RQ4 evaluates the sensitivity of SpecGuard to controlled, seeded defects. The empirical study uses a corpus of 64 requirements from the CVA6 RISC-V specification [21]. CVA6 is an open-source 64-bit RISC-V processor core developed by the OpenHW Group. It is used here as an industrial-grade representative open-source FPGA specification of comparable complexity, not as a DO-254-certified product. The single WARN requirement (L1W-60) contains a vagueness smell triggered by the quantifier "some". The two FAIL requirements (PPA-50 and PPA-60) contain explicit "TBD" placeholders and receive high-severity completeness penalties. Because their texts contain neither a modal commitment nor a measurable target, verifiability remains at the 0.50 baseline; the placeholder itself does not penalize verifiability. The detected smells are concentrated in five types. Table 3 shows the distribution.

Table 2. Detected Smell Distribution in the CVA6 Corpus

Smell type	**Count**
implicit_reference	4
negative_statement	3
missing_unit	2
placeholder	2
vagueness	1

V. RESULTS

1. RQ1: Smell Distribution

SpecGuard detected 12 smell events across 11 of the 64 requirements (17.2% flag rate); the event count exceeds the requirement count because one requirement carries two distinct smells. This corresponds to a relatively low concentration of detected smells, which is expected for a mature open-source specification. Table 2 summarises the corpus-level results.

Table 3. Corpus-Level Specguard Results On Cva6

Indicator	Value
Requirements	64
Req. with smells	11
Smell events	12
PASS	61
WARN	1
FAIL	2
Avg. overall	0.888
Avg. verifiability	0.777
Avg. completeness	0.961

Six of the eleven smell types in the SpecGuard catalog were not triggered on the CVA6 corpus. This pattern is consistent with the description of CVA6 as a mature open-source specification: ambiguity, subjectivity, comparability, optionality, weakness, and non-verifiability are typically associated with under-specified or stylistically weak requirements, which are less common in projects that have undergone multiple rounds of community review.

2. RQ2: Linguistic Metrics

The linguistic metrics show a profile consistent with technical requirements: moderate readability, relatively short requirements, and moderate syntactic complexity. Table 4 reports the descriptive statistics.

Table 4. Descriptive Statistics Of Linguistic Metrics

Metric	Mean	Median	Max
FRE	46.04	47.60	78.87
FKGL	10.84	10.10	30.97

Metric	Mean	Median	Max
MDL	3.06	2.89	5.88
Max dep.	16.41	12.00	91.00
Tokens	22.20	19.50	101.00
Lex. density	0.40	0.40	0.60

The average Flesch Reading Ease score of 46.04 falls within the range expected for technical specifications. The average FKGL value of 10.84 is lower than might be expected for complex engineering prose, which the short, structured form of many CVA6 requirements can help explain. The average MDL of 3.06 indicates moderate syntactic complexity.

3. RQ3: Complementarity of Framework Layers

Correlation analysis was used to examine whether different layers of the framework provide complementary information. Table 5 reports selected Spearman correlations between SpecGuard indicators.

Table 5. Spearman Correlations Between Selected Indicators

Pair	Coeff.	Interpretation
Smell count − overall	−0.531	strong neg.
Smell count − verifiability	−0.474	moderate neg.
MDL − overall	+0.447	moderate pos.
MDL − verifiability	+0.420	moderate pos.
FRE − overall	+0.215	weak pos.
FKGL − overall	−0.073	negligible
Token count − MDL	+0.598	strong pos.

The weak or negligible relationship between Flesch-based metrics and composite scores supports the use of multiple layers. Readability formulas alone are insufficient for assessing the quality of technical FPGA requirements. The positive correlation between MDL and overall score is counterintuitive at first glance. Still plausible: in the CVA6 corpus, longer requirements with higher MDL tend to contain more explicit quantification, protocol references, and standard citations, thereby improving their verifiability.

4. RQ4: Seeded-Fault Sensitivity Check

A seeded-fault experiment was conducted to evaluate the detectors' sensitivity to controlled, rule-detectable defects. For each of the five mutation types (ambiguity, vagueness, optionality, placeholder, and comparative forms without a baseline), ten PASS requirements were sampled and mutated, producing 50 seeded-defect cases.

Across these 50 controlled mutations, every injected defect was detected (10/10 per type). This result should be interpreted as a sanity check of detector behavior, not as a general precision or recall estimate for industrial requirements. The experiment does not evaluate semantic defects or false-positive rates on independently annotated corpora.

Because these mutations are drawn from the detector's own trigger lexicon, the 50/50 result (100%; 10/10 per type) is expected by construction and is only an end-to-end implementation check. Two additional checks probe coverage. None of the 50 mutations using expressions verified as disjoint from the same-category trigger vocabulary was detected (0/50, 0%). An LLM-authored blind set yielded 14/49 detections (28.6%). Thus, the closed catalog does not generalize reliably to unseen forms. The repository also reports an exploratory estimate of 8/64 (12.5%) clean-corpus flag after an author-defined exclusion of three known defects; without independent expert ground truth, it is not a formal false-positive rate. Generalization and precision require annotated data, expert calibration, lexicon expansion, or complementary semantic analysis independently.

5. Worked Example: From Detected Smell to Gate Decision

To illustrate how an individual smell propagates to the gate decision, we trace two CVA6 requirements through the scoring pipeline.

Requirement L1W-60 — "Some physical memory regions shall be configurable as not L1WTD cacheable at design time." — triggers a single vagueness smell on the quantifier "some" (high severity). Vagueness is a verifiability-related smell, so the penalty is applied to the verifiability sub-score, which drops to 0.10, while completeness and consistency remain at 1.00. The overall score is $0.30 \cdot 1.00 + 0.25 \cdot 1.00 + 0.45 \cdot 0.10 = 0.595$, which falls in the WARN band [0.50, 0.75). The requirement is structurally complete but not yet objectively testable, so it is flagged for review rather than rejected.

Requirement PPA-50 — "TBD: Placeholder for single-precision floating performance per MHz." — triggers a single placeholder smell on "TBD" (high severity). Completeness starts at 0.80 for this text and receives both the ordinary high-severity penalty and the additional 1.5 × placeholder penalty, yielding 0.00 after bounding. The placeholder does not directly penalize verifiability; the score remains at its 0.50 baseline because the text has neither a modal commitment nor a measurable target. Consistency remains 1.00. Overall is $0.30 \cdot 0.00 + 0.25 \cdot 1.00 + 0.45 \cdot 0.50 = 0.475$, below the 0.50 FAIL threshold. The requirement must be completed before HDL design or verification.

These cases show that outcomes depend on both text-derived baselines and the implemented mappings (Table 6): vagueness reduces verifiability in L1W-60, whereas the placeholder reduces completeness in PPA-50. SpecGuard emits the smell, trigger, severity, sub-scores, and gate decision as JSON for review and reproducibility.

Table 6. Worked Example: Smell-To-Score Propagation

Req.	**Smell (trigger)**	**Sev.**	**Comp.**	**Cons.**	**Verif.**	**Overall**	**Gate**
L1W-60	vagueness ("some")	High	1.00	1.00	0.10	0.595	WARN
PPA-50	placeholder ("TBD")	High	0.00	1.00	0.50	0.475	FAIL

VI. DISCUSSION AND THREATS TO VALIDITY

The pilot study indicates that SpecGuard can be applied to a mature open-source FPGA-related specification without rewriting the requirements into a controlled template language. The results also support the design decision to preserve both diagnostic details and aggregate indicators. A single score would not explain whether a requirement is problematic because of vagueness, missing units, weak modality, or syntactic complexity. In practical terms, SpecGuard is intended for the requirements author or verification engineer at the point before requirements enter HDL design and assertion development. It acts as an automated pre-review gate in the requirements-capture phase of a DO-254 lifecycle: the engineer runs it on a draft specification, receives a per-requirement smell report and a PASS/WARN/FAIL verdict, and resolves flagged issues before the requirements are baselined. The practical value is reduced downstream rework, since vague, unverifiable, or incomplete requirements are caught at authoring time rather than during a Stage-of-Involvement review or test development, when correction is substantially more expensive.

The study does not demonstrate certification readiness, nor does it claim that CVA6 is DO-254 compliant. CVA6 is used as a representative open-source artifact of comparable hardware complexity.

Construct validity is limited because requirements smells are only proxies for quality problems. A smell-free requirement may still be incorrect, incomplete, or inconsistent with stakeholder intent, and no independent expert annotation is available as ground truth. The distribution of smells is therefore an output of the current operational definitions, not a validated defect distribution. Lexicon coverage is limited: recall was 0/50 (0%) on the deliberately disjoint tier and 14/49 (28.6%) on the LLM-authored blind tier, versus 50/50 on the own-lexicon sanity check. The first measures the closed-vocabulary boundary; the second is less constrained but not expert-authored. Neither is a validated industrial precision/recall estimate.

External validity is limited by the single-corpus design. Results may differ for other FPGA projects such as Ibex, OpenTitan, CV32E40P, or proprietary industrial specifications. Conclusion validity is also limited because the study is descriptive and exploratory. Correlation values and seeded-fault results should not be interpreted as statistically generalizable claims. A further limitation on construct validity is that the current smell catalog does not distinguish among requirement abstraction levels. In a hierarchical specification, high-level (parent) requirements are expected to be more abstract and to tolerate a degree of ambiguity that is inappropriate at the low-level (child) requirement layer, where requirements should be concrete and directly verifiable. SpecGuard currently applies a single rule set to all requirements, regardless of level; consequently, the reported flag rate may overstate the number of quality issues for intentionally high-level requirements. Level-aware smell thresholds, relaxing ambiguity penalties for parent requirements while tightening verifiability penalties for child requirements, are identified as future work.

VII. CONCLUSION

This paper presented SpecGuard, a multi-layer framework for assessing the quality of natural-language FPGA requirements. The framework combines deterministic requirements-smell detection, linguistic metrics, composite quality scores, and PASS/WARN/FAIL gate decisions. A pilot study on 64 CVA6 RISC-V requirements demonstrated that the framework can process an industrial open-source FPGA-related specification and produce diagnostic outputs suitable for further review. The main contribution is not a claim of certification readiness, but a reproducible quality-assessment baseline for natural-language FPGA specifications. Future work will extend the evaluation to additional open-source hardware corpora, introduce expert annotations, empirically calibrate severity weights, and investigate graph-based traceability and compliance analysis.

ACKNOWLEDGEMENT

The author thanks Prof. Yevhen Brezhniev (KhAI) for scientific supervision and Prof. Vyacheslav Kharchenko for methodological guidance throughout this research.

REFERENCES

[1] ISO/IEC/IEEE 29148:2018, *Systems and software engineering — Life cycle processes — Requirements engineering*. Geneva: ISO/IEC/IEEE, 2018.

[2] RTCA DO-254, *Design Assurance Guidance for Airborne Electronic Hardware*. Washington, DC: RTCA, Inc., 2000.

[3] RTCA DO-178C, *Software Considerations in Airborne Systems and Equipment Certification*. Washington, DC: RTCA, Inc., 2011.



[4] L. Zhao et al., "Natural language processing for requirements engineering: A systematic mapping study," *ACM Computing Surveys*, vol. 54, no. 3, article 55, 2021.

[5] R. Sonbol, G. Rebdawi, and N. Ghneim, "The use of NLP-based text representation techniques to support requirement engineering tasks: A systematic mapping study," *IEEE Access*, vol. 10, pp. 62811–62830, 2022.

[6] L. Montgomery, D. Fucci, A. Bouraffa, L. Scholz, and W. Maalej, "Empirical research on requirements quality: A systematic mapping study," *Requirements Engineering*, vol. 27, no. 2, pp. 183–209, 2022.

[7] H. Femmer, D. Mendez Fernandez, S. Wagner, and S. Eder, "Rapid quality assurance with requirements smells," *Journal of Systems and Software*, vol. 123, pp. 190–213, 2017.

[8] A. Veizaga, S. Y. Shin, and L. C. Briand, "Automated smell detection and recommendation in natural language requirements," *IEEE Transactions on Software Engineering*, vol. 50, no. 4, pp. 695–720, 2024.

[9] E. Gentili and D. Falessi, "Characterizing requirements smells," in *PROFES 2023*, LNCS, vol. 14483, Cham: Springer, 2024, pp. 387–398.

[10] T. Wang, W. Wang, and Y. Yang, "AirReq: Automated requirements smell detection and elimination for commercial aircraft systems," in *IEEE Requirements Engineering Conference Workshops*, 2025, pp. 173–180.

[11] C. Kumar Sahu, V. Rai, and V. Roshan, "Well-formed quality of system requirements for verifying to ISO 29148-2018: A natural language processing (NLP) based framework and quantitative metric," in *Proc. ASME IDETC-CIE*, 2024, paper V02BT02A013, doi: 10.1115/DETC2024-139583.

[12] M. Zakeri-Nasrabadi and S. Parsa, "Natural language requirements testability measurement based on requirement smells," *Neural Computing and Applications*, vol. 36, no. 12, pp. 6435–6458, 2024.

[13] L. d. M. Barbosa, I. C. A. de Oliveira, C. S. Cerqueira, and A. E. C. da Cunha, "Mean dependency length: A new metric for requirements quality," *INCOSE International Symposium*, vol. 34, no. 1, pp. 1021–1035, 2024, doi: 10.1002/iis2.13193.

[14] K. Großer, J. Frattini, J. Lehmann, and D. Mendez, "Templated requirements: A structured catalog and an evaluation of usage," *Requirements Engineering*, vol. 29, no. 3, pp. 401–428, 2024.

[15] J. Marques, S. Yelisetty, L. A. V. Dias, and A. M. Cunha, "Software product assurance for IEC 62279, IEC 62304, and DO-178C: A literature review," in *ITNG 2019*, pp. 615–620.

[16] O. Kosenkov, M. Borg, D. Mendez, and J. Münch, "Vision for an artifact-based approach to regulatory requirements engineering," *Information and Software Technology*, vol. 172, article 107477, 2024.

[17] A. Menon et al., "Enhancing large language models for hardware verification: A novel SystemVerilog assertion dataset," *ACM Transactions on Design Automation of Electronic Systems*, 2025.

[18] J. Kuchta, *Completeness and consistency of the system requirement specification*, in Position Papers of the 2016 Federated Conf. on Computer Science and Information Systems (FedCSIS), ACSIS vol. 9, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., 2016, pp. 265-269, doi: 10.15439/2016F468.

[19] M. Honnibal and I. Montani, "spaCy: Industrial-strength natural language processing in Python," 2017. [Online]. Available: https://spacy.io.

[20] *textstat*: A Python package for calculating readability and complexity statistics. [Online]. Available: https://github.com/textstat/textstat.

[21] OpenHW Group, *CVA6 Requirements Specification*, Revision 1.0.1. [Online]. Available: https://docs.openhwgroup.org/projects/cva6-user-manual/02_cva6_requirements/cva6_requirements_specification.html.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Counterfactual Explanations for Gallstone Prediction: A Comparative Evaluation of DiCE and PertCF

Busra Duygu Çelik
Graduate School of Natural and Applied Sciences
Gazi University
Ankara, Türkiye
busraduygucelik@gazi.edu.tr
ORCID:0000-0001-9640-1025

Semiha Güngör
Hacı Bayram Veli University
Ankara, Türkiye
semiha.gungor@hbv.edu.tr
ORCID: 0000-0002-8862-1340

Adem Tekerek
Department of Computer Engineering, Gazi University
Ankara, Türkiye
atekerek@gazi.edu.tr
ORCID:0000-0002-0880-7955

Abstract
Gallstone disease is a common gastrointestinal disorder for which non-invasive pre-imaging risk prediction using routine clinical and bioimpedance data may serve as a low-cost screening aid. Beyond predictive performance, counterfactual explanations can show clinicians how changes in input features could alter a model's decision, without implying that such changes represent actionable clinical recommendations. However, the consistency and trade-offs of different counterfactual generation methods remain insufficiently characterized in gallstone prediction. This study compares two counterfactual explanation methods, DiCE and PertCF, using 31 test-set patients predicted as gallstone-positive by a tree-based ensemble trained on clinical and bioimpedance data from 319 patients. The methods were evaluated using matched, paired statistical tests under a common backward-elimination post-processing protocol. DiCE and PertCF did not differ significantly in sparsity (median 2.0 features for both, Holm-adjusted p=0.313) or proximity (Holm-adjusted p=0.544). In contrast, PertCF was substantially faster (median 0.625 s vs. 3.80 s; Holm-adjusted $p < 0.001$). PertCF also exhibited a numerically higher rate of one-sided changes within strongly correlated feature pairs than DiCE (54.8% vs. 38.7%), although this difference was not statistically significant (exact McNemar p=0.125). The ensemble classification backbone showed no significant advantage over a simpler two-model average under nested cross-validation (p=0.206). SHAP analysis and univariate comparisons provided convergent evidence highlighting C-Reactive Protein and Vitamin D. Overall, neither counterfactual method was uniformly superior: PertCF demonstrated a clear computational advantage, while differences in the preservation of strong empirical feature dependencies remained statistically inconclusive.

Keywords: counterfactual explanations, DiCE, PertCF, explainable AI, gallstone disease, clinical decision support

I. INTRODUCTION

Gallstone disease (cholelithiasis) is a common gastrointestinal disorder worldwide, with prevalence varying across populations, sex, and age [1]. Although many gallstones remain asymptomatic, symptomatic disease may progress to biliary colic, cholecystitis, or pancreatitis and may ultimately require surgical management [2]. Abdominal ultrasonography is the standard first-line imaging modality for suspected gallstone disease [2]. Non-invasive risk prediction using routinely available clinical, laboratory, and bioimpedance measurements may therefore provide a low-cost pre-imaging screening aid. A publicly available dataset comprising 319 individuals with clinical, laboratory, and bioimpedance measurements [3] has enabled several machine-learning studies of gallstone prediction. Across these studies, C-reactive protein (CRP) and Vitamin D have repeatedly emerged among prominent predictive features [3-6]. Existing work has also incorporated post-hoc explanation methods such as SHAP and LIME, while DiCE has been used to generate individual counterfactual explanations [5]. However, to our knowledge, existing studies on this benchmark have not systematically compared multiple

counterfactual generation methods using paired measures of coverage, sparsity, proximity, computational cost, and preservation of empirical feature dependencies. Beyond predictive performance, counterfactual explanations provide instance-level information about how changes in the input space could alter a model's prediction, complementing feature-attribution methods that primarily characterize feature contributions [7]. This study considers two structurally distinct approaches: DiCE [8], implemented with its genetic search strategy, and PertCF [9], a perturbation- and ablation-based approach designed for computational efficiency. Previous counterfactual benchmarking studies have shown that generation strategies can differ substantially in validity, proximity, sparsity, and computational cost [10]; however, these trade-offs have not been systematically characterized for the gallstone prediction benchmark considered here. Accordingly, the primary objective of this study is to conduct a paired comparison between DiCE and PertCF on the same patient cohort using a common backward-elimination post-processing protocol. Both methods are evaluated against the same fixed tree-based classification backbone, thereby isolating differences attributable to the counterfactual generation procedure. The classification backbone is assessed separately using nested cross-validation to determine whether its optimized weighting provides a measurable benefit over a simpler ensemble strategy. The main contributions are:

- A matched, paired evaluation of DiCE and PertCF in terms of coverage, sparsity, proximity, and computational cost; paired differences in sparsity, proximity, and computational cost were assessed using Holm-adjusted Wilcoxon signed-rank tests.
- A dependency-consistency analysis quantifying one-sided counterfactual changes within strongly correlated feature pairs ($|r| \geq 0.80$) identified from the training data, with paired differences evaluated using an exact McNemar test.
- A complementary analysis examining convergence among univariate associations, SHAP-based feature attribution, and features modified by counterfactual explanations.

II. RELATED WORK

Esen et al. [3] compiled the dataset used here and reported 85.4% accuracy, with Vitamin D, CRP, total body water, and lean mass as the top predictors. Kothari et al. [4] compared a tuned XGBoost against RF and AdaBoost (test AUC=0.859) using SHAP and LIME, without repeated cross-validation or significance testing. Das et al. [5] proposed an SLOA-optimized CatBoost (5-fold CV accuracy 80.4%) using SHAP, LIME, and DiCE, again finding CRP and Vitamin D dominant. Akben and Yumrutas [6] used a stacked SVM/logistic regression/decision tree ensemble with nested cross-validation (82.3% accuracy, 85.7% AUC) and explicitly reported that strong multicollinearity (VIF > 100) destabilizes Vitamin D's SHAP direction, an independent parallel to the dependency-consistency issue documented in Section IV. None of these studies benchmarks the quality of counterfactual outputs or compares multiple counterfactual methods. Counterfactual explanations. Wachter et al. [7] introduced the foundational framework. DiCE [8] uses genetic search to generate diverse counterfactuals; PertCF [9] uses ablation-based feature importance with greedy median shifting at a much lower cost. Bayrak et al. [10] conducted a comparative benchmark of multiple counterfactual explanation methods across several classifiers in network-resilience classification, demonstrating that counterfactual methods can differ substantially in output quality and computational efficiency. The present study extends this comparative perspective to clinical tabular data by directly evaluating DiCE and PertCF on the gallstone prediction task.

Calibration and clinical utility. Calibration provides an additional, secondary check on a classifier's clinical relevance beyond discrimination [11]; this has not previously been reported for this dataset, though a full decision-analytic evaluation is beyond the scope of this paper.

III. METHODOLOGY

1. Dataset

The public gallstone dataset [3] contains 319 patients with 38 clinical and bioimpedance features and a binary gallstone-presence label: 161 negative (50.5%) and 158 positive (49.5%). Data were split 80/20 (stratified; n=255 train, n=64 test).

2. Preprocessing and Descriptive Comparison

Group comparisons (Table 1) used only the training set (n=255) and are strictly descriptive; they were not used for feature selection or model development, and all 38 features were entered into every model unchanged. Normality was assessed with the Shapiro-Wilk test; Welch's t-test or Mann-Whitney U was used accordingly, and chi-square/Fisher's exact test for categorical variables, with Benjamini-Hochberg FDR correction [12] across 38 features.

Table 1. Selected Features with The Strongest Group Differences (N=255)

Feature	**Test**	**p_FDR**	**Effect size**
C-Reactive Protein (CRP)	Mann-Whitney U	6.70e-19	r=0.582
Vitamin D	Mann-Whitney U	2.59e-10	r=-0.438
Aspartate Aminotransferase (AST)	Mann-Whitney U	3.01e-6	r=-0.334
BMI, Height, ALP	Welch t / MWU	n.s. (>0.05)	small

Table 2. Strongest Correlated Feature Pairs (Top 6 Of 19, |R|>=0.80)

Feature 1	**Feature 2**	r
Total Body Fat Ratio	Lean Mass	-0.998
Total Body Water	Muscle Mass	0.917
BMI	Total Fat Content	0.910
Total Body Water	Extracellular Water	0.897
Intracellular Water	Muscle Mass	0.888
Total Fat Content	Visceral Fat Area	0.885

3. Classification Models

Four tree-based models were compared: CatBoost [13], XGBoost [14], LightGBM [15], and Random Forest [16]. Hyperparameters were tuned with Optuna [17] (60 trials each, 5-fold CV, AUC objective) on the training set only.

4. AUC-Optimized Weighted Ensemble

The four models were combined by soft voting with weights optimized via differential evolution [18] on train-only out-of-fold predictions, maximizing AUC alone (no diversity objective). The single-run weights were CatBoost=0.485, XGBoost=0.497, LightGBM=0.000, RF=0.018. Their stability and the intercorrelation among the base models are assessed in Section IV-3.

5. Statistical Validation

Repeated stratified 5-fold x 10-repeat cross-validation was used for all models, with hyperparameters fixed as in Section III-C (base hyperparameters were selected once on the training set and were not re-tuned within each fold; only the ensemble weights were nested, below). Repeated-CV estimates should therefore not be interpreted as fully nested estimates of the complete model-selection pipeline; a full nested hyperparameter search was not computationally feasible within this conference paper timeline. A dedicated nested procedure tested whether the single-run weight optimization was optimistic: for each of 50 outer folds, weights were re-optimized on an inner 5-fold out-of-fold split of that fold's training portion only, the models were refitted on the full outer training data, and the models were evaluated on the held-out outer test fold. Because the 50 folds are not independent (the same 255 patients recur across repeats), confidence intervals were calculated using the Nadeau-Bengio correction [19], which adjusts the variance estimator for the train/test size ratio; the same correction was used for a paired t-test of the 4-model vs. 2-model difference on identical folds. McNemar's test [20] (accuracy) and DeLong's test [21] (AUC), with Holm-Bonferroni correction, assessed the ensemble's superiority over each base model on the held-out test set.

6. Calibration

Calibration intercept and slope were estimated by logistic recalibration of each model's logit-transformed probabilities against the outcomes, along with the Brier score, providing a secondary check on the classifier backbone's clinical relevance beyond discrimination [11].

7. Explainable AI

SHAP [22] values were computed per model using TreeExplainer as a secondary, supporting analysis; cross-model agreement was assessed using Spearman/Kendall correlations and top-5/top-10 Jaccard overlap.

The primary analysis compares two counterfactual methods: DiCE [8] (genetic search) and PertCF [9] (ablation-based greedy median-shifting), both run on all 31 test patients predicted gallstone-positive by the ensemble, with age, sex, height, and diagnosed comorbidities held constant. Laboratory values (e.g., CRP, Vitamin D) were left as mutable target biomarkers for the model's decision boundary, not as a prescription for direct clinical intervention; this distinction is revisited in Section V-3. DiCE's own validity claims were never trusted directly; every raw candidate was independently re-verified against the classifier. Both methods' raw outputs were passed through an identical backward-elimination step, reverting changes to features one at a time with re-verification after each step. Hence, the comparison reported below is DiCE+BE vs. PertCF+BE, not the raw outputs. Because DiCE can yield up to three candidates per patient, backward elimination was applied to each of the (up to three)

raw candidates, and only the resulting lowest post-elimination proximity candidate per patient was retained, yielding a genuinely matched 31-vs-31 sample for all paired tests. Metrics: coverage, sparsity, normalized-L1 proximity, and range feasibility (changed values within the observed training range). Each generated counterfactual was further checked against the 19 correlated pairs from Section III-2: changing exactly one member of a pair without the other is flagged as a dependency-consistency violation, a proxy for, not proof of, physiological implausibility, since a high correlation does not imply a deterministic per-patient dependency. Sparsity/proximity/runtime differences were tested with paired Wilcoxon signed-rank tests (Holm-Bonferroni corrected, matched-pairs rank-biserial effect size); the violation-rate difference was tested with matched McNemar.

IV. RESULTS

1. Classification Performance

Repeated CV AUC ranged 0.838-0.856 across base models; the naive (single-run) ensemble reached 0.863 (Table 3). On the held-out test set (n=64, Table 4), XGBoost outperformed the ensemble (90.6% vs. 85.9% accuracy) and also showed the most favorable calibration indices. This discrepancy with the CV ranking likely reflects small-sample test-set variability (n=64) rather than true model superiority; Section IV-3 addresses why a single fixed backbone was nonetheless used for the counterfactual comparison.

Table 3. Classification Performance (Repeated CV and Nested)

Model	Accuracy (%)	AUC
CatBoost	77.61	0.8559
XGBoost	77.41	0.8528
LightGBM	76.94	0.8422
Random Forest	75.92	0.8375
Ensemble (naive, single-run weights)	78.39	0.8628
Ensemble (nested, honest estimate)	77.57	0.8519
2-model average, CatBoost+XGBoost (nested)	78.35	0.8628

Table 4. Held-Out Test Set Results (N=64)

Model	Accuracy (%)	AUC	TN	FP	FN	TP
CatBoost	76.56	0.883	26	6	9	23
XGBoost	90.62	0.926	29	3	3	29
LightGBM	82.81	0.909	29	3	8	24
Random Forest	78.12	0.875	27	5	9	23
Ensemble	85.94	0.910	28	4	5	27

2. Statistical Significance vs. Base Models

After Holm-Bonferroni correction, the ensemble showed no significant superiority over any base model (Table 5, all p_Holm>0.05). This does not prove the absence of a true difference; given n=64, both limited power and a genuinely small/negligible difference remain plausible and cannot be distinguished here.

Table 5. Ensemble Vs. Base Models (Holm-Bonferroni Corrected)

Comparison	McNemar p_Holm	DeLong p_Holm
Ensemble vs CatBoost	0.125	0.085

Ensemble vs XGBoost	0.539	0.259
Ensemble vs LightGBM	0.688	0.938
Ensemble vs Random Forest	0.539	0.148

3. Ensemble Weight Instability and Nested Evaluation

Base-model predictions were highly correlated (Pearson r=0.90-0.94), and under nested evaluation (Section III-5), the optimized four-model ensemble did not significantly outperform the simple CatBoost-XGBoost average (AUC 0.852 vs. 0.863; Nadeau-Bengio corrected paired test, p=0.206). Ensemble weighting, therefore, provided no statistically established benefit for this dataset size. The ensemble was nevertheless retained as the pre-specified fixed backbone for the counterfactual comparison, since both DiCE and PertCF were evaluated against the same, unchanged backbone throughout; changing it post hoc based on a held-out test split would itself constitute the kind of small-sample overfitting this analysis warns against.

4. Calibration

XGBoost showed the most favorable calibration indices on the held-out sample (slope=0.953, Brier=0.098); all models showed a positive calibration intercept (0.107-0.207) in this test set, suggesting underprediction of absolute risk that may require external validation to generalize. Full calibration and decision-curve results are omitted here for brevity, as they are secondary to this paper's counterfactual comparison; the AUC values already summarize classification discrimination in Table 4.

5. Counterfactual Comparison: DiCE vs. PertCF

Both methods reached 100% coverage and 100% range feasibility (every changed value fell within the observed training range) on all 31 patients. DiCE's raw (pre-backward-elimination) output changed 27.8 features on average; backward elimination reduced this by 88.9% to 3.09. PertCF's raw output changed 8.90 features, reduced by 69.6% to 2.71 (Table 6), indicating that DiCE's raw genetic-search solutions contained substantially more removable feature changes under the common backward-elimination procedure.

Table 6. Raw Vs. Backward-Eliminated Sparsity

Method	**Raw sparsity**	**After BE**	**Reduction**
DiCE	27.78	3.09	88.9%
PertCF	8.90	2.71	69.6%

Table 7. Paired Wilcoxon Test (N=31)

Dimension	**DiCE (median)**	**PertCF (median)**	**p_Holm**	**Effect size**
Sparsity	2.0	2.0	0.313	-0.38
Proximity	0.932	0.977	0.544	0.127
Runtime (s)	3.80	0.625	<0.001	-1.00

Dependency-consistency (Table 8) showed PertCF violating at least one correlated pair in 54.8% of matched counterfactuals vs. 38.7% for DiCE (mean 2.74 vs. 1.71 violations per case); this difference was not significant (exact McNemar p=0.125, 7 discordant pairs). This is the primary result used for the dependency-consistency comparison.

Table 8. Dependency-Consistency Violations (Matched, N=31 Each)

Method	**Violation rate**	**Mean violations/case**
DiCE (best per-patient candidate)	38.7%	1.71

PertCF	54.8%	2.74

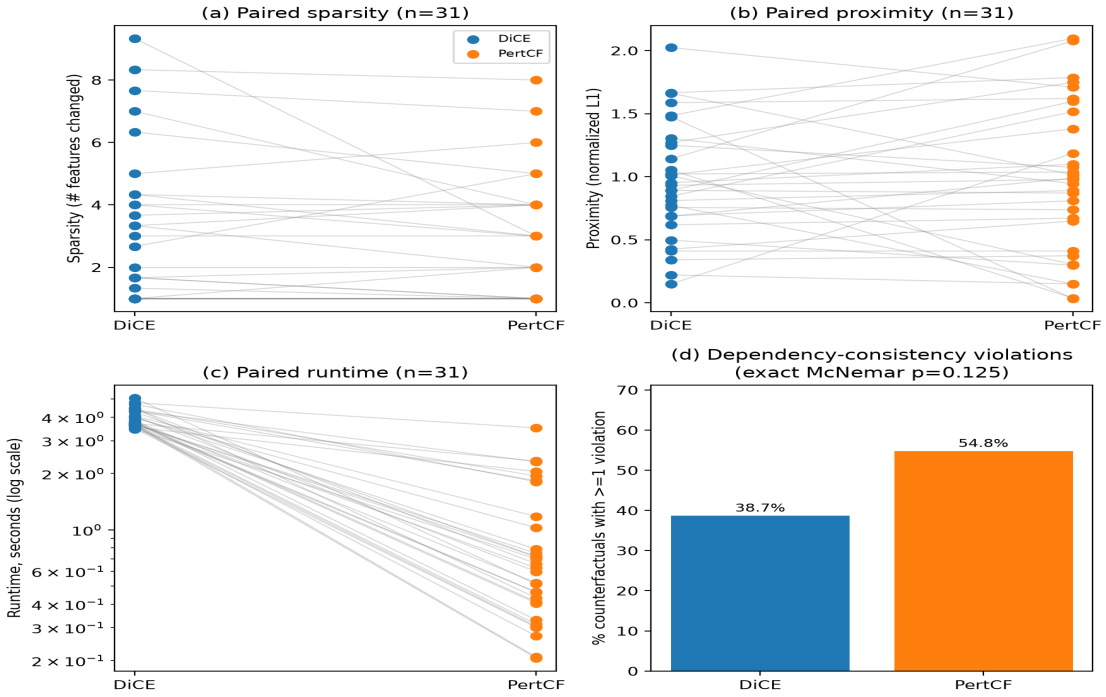


Figure 1. DiCE vs. PertCF, matched comparison: (a) sparsity, (b) proximity, (c) runtime (log scale), (d) dependency-consistency violation rate.

6. SHAP Cross-Model Agreement

SHAP values from the four models consistently ranked CRP and Vitamin D highest, associated with the model's output rather than implying a causal risk-increasing or risk-decreasing relationship. This convergence with Sections IV-B/E should be read as cross-method convergence rather than independent validation, since all four models share the same training data and yield highly correlated predictions. Given the strong correlations documented in Table 2, SHAP attribution may also be split across correlated body-composition features rather than assigned entirely to a single variable, a known limitation of Shapley-based attribution in the presence of multicollinearity. Cross-model SHAP rankings showed high agreement (Spearman rho=0.849-0.962), with CRP and Vitamin D consistently among the top-ranked variables across all four models.

V. DISCUSSION

1. Interpreting the Counterfactual Comparison

The similar post-elimination sparsity despite markedly different raw search behavior (Table 6) suggests that the common backward-elimination stage, rather than the native generator alone, substantially shapes the final counterfactual representation. DiCE's genetic search and PertCF's greedy ablation converge toward comparably sparse solutions once redundant changes are removed, even though their raw outputs differ by nearly an order of magnitude. This indicates partial agreement between the two strategies after post-processing, rather than proof that either has found the model's "true" decision boundary. PertCF's speed advantage converges with, though does not directly replicate, findings in a different domain [10]. No counterfactual method was uniformly superior

across all evaluated dimensions: PertCF was significantly faster, sparsity/proximity were statistically indistinguishable, and PertCF's numerically higher rate of dependency-consistency violations did not reach significance (Section IV-5).

2. Ensemble Finding

The ensemble backbone's modest CV advantage did not hold under nested evaluation or on the test set. High base-model correlation (0.90-0.94) plausibly contributes to this instability, though sample size and optimizer variance may also play a role. This is presented as an honest methodological finding, not as a limitation of the counterfactual comparison, which used a single fixed backbone throughout.

3. Clinical Interpretation and Limitations

Counterfactual outputs are model-derived changes to features, not causal or actionable clinical recommendations. CRP, for instance, is an acute-phase reactant rather than a directly controllable input. Calibration and decision-curve results are exploratory, single-center, and require external validation before any claim of clinical utility can be made. Because DiCE generated up to three candidates per patient and only the lowest-proximity one was retained, whereas PertCF produced a single deterministic solution, the paired comparison may be conservative with respect to PertCF on proximity; this selection rule was a deliberate design choice, disclosed here for transparency. Because DiCE uses unseeded stochastic (genetic) search, its output and any significance test built on it may vary between runs; future studies should quantify this run-to-run variability across multiple independent runs rather than relying on a single run, as done here for the main analysis. More broadly, dependency-consistency violations are a proxy for, not proof of, physiological implausibility, and neither method enforces joint updates for correlated features, a concrete direction for future work.

VI. CONCLUSION

This study compared DiCE and PertCF for gallstone risk prediction using matched, paired statistical tests. Neither method was uniformly superior: sparsity and proximity were statistically indistinguishable, PertCF was significantly faster, and PertCF showed a higher but non-significant rate of dependency-consistency violations. The ensemble classification backbone showed no significant advantage over a simple two-model average under nested cross-validation, consistent with high inter-model correlation. SHAP and univariate analysis consistently identified CRP and Vitamin D as the strongest predictors. Future work should develop counterfactual generation methods that jointly update correlated features and should quantify DiCE's run-to-run variability directly.

REFERENCES

[1]. X. Wang et al., "Global epidemiology of gallstones in the 21st century: A systematic review and meta-analysis," Clinical Gastroenterology and Hepatology, vol. 22, no. 8, pp. 1586–1595, 2024.

[2]. H. Patel and J. Jepsen, "Gallstone disease: Common questions and answers," American Family Physician, vol. 109, no. 6, pp. 518–524, 2024.

[3]. Esen, H. Arslan, S. Akturk, Esen, M. Gulsen, N. Kultekin, and O. Ozdemir, "Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data," Medicine (Baltimore), vol. 103, no. 8, p. e37258, 2024.

[4]. A. Kothari, U. Salvi, A. Aylani, and S. Shrivastava, "An interpretable machine learning framework for gallstone disease prediction using XGBoost and explainable AI," Procedia Computer Science, vol. 283, pp. 2455–2466, 2026.

[5]. P. Das, M. A. U. Khan, P. Sarker, and A.-A. Nahid, "Gallstone disease classification using SLOA-optimized CatBoost classifier with explainable AI," PLoS One, vol. 21, no. 6, art. e0342945, 2026, doi: 10.1371/journal.pone.0342945.

[6]. S. B. Akben and H. Yumrutas, "Interpretable machine learning for gallstone risk prediction: An ensemble stacking approach with SHAP analysis," Computer Methods in Biomechanics and Biomedical Engineering, 2026, doi: 10.1080/10255842.2026.2714840.

[7]. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box," Harvard J. Law Technol., vol. 31, no. 2, pp. 841–887, 2017.

[8]. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in Proc. ACM Conf. Fairness, Accountability, Transparency, 2020, pp. 607–617.

[9]. B. Bayrak and K. Bach, "PERTCF: A perturbation-based counterfactual generation approach," in Proc. Int. Conf. Innov. Techn. Appl. Artif. Intell., 2023, pp. 174–187.

[10]. B. Bayrak, M. Gullu, S. N. Karahan, and B. D. Celik, "Benchmarking counterfactual explanation methods for network resilience classification under seismic stress," IEEE Access, vol. 14, pp. 99586–99605, 2026.

[11]. B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," BMC Medicine, vol. 17, art. 230, 2019.

[12]. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate," J. Roy. Stat. Soc. Series B, vol. 57, no. 1, pp. 289–300, 1995.

[13]. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Proc. 32nd Int. Conf. Neural Inf. Process. Syst., 2018, pp. 6639–6649.

[14]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[15]. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 3149–3157.

[16]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[17]. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2019, pp. 2623–2631.

[18]. R. Storn and K. Price, "Differential evolution," J. Global Optimization, vol. 11, no. 4, pp. 341–359, 1997.

[19]. C. Nadeau and Y. Bengio, "Inference for the generalization error," Machine Learning, vol. 52, no. 3, pp. 239–281, 2003.

[20]. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.

[21]. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated ROC curves," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.

[22]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 4765–4774.

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

Classification of Eight Different Retinal Conditions in Retinal OCT Images Using Convolutional Neural Networks

Semiha Güngör
Hacı Bayram Veli University
Ankara, Türkiye
semiha.gungor@hbv.edu.tr
ORCID: 0000-0002-8862-1340

Busra Duygu Çelik
Gazi University
Ankara, Türkiye
busraduygucelik@gazi.edu.tr
ORCID:0000-0001-9640-1025

Adem Tekerek
Gazi University
Ankara, Türkiye
atekerek@gazi.edu.tr
ORCID: 0000-0002-0880-7955

Abstract
In this study, a convolutional neural network-based approach was developed to automatically classify eight retinal conditions from retinal optical coherence tomography (OCT) images. The balanced dataset comprising AMD, CNV, CSR, DME, DR, DRUSEN, MH, and NORMAL classes is divided into training, validation, and test sets in an 80%/10%/10 % split. Images were resized to 224×224 pixels, and the CNN model, consisting of three convolutional blocks, was trained for 10 epochs. The model achieved 98.68% accuracy on the training set, 81.14% on the validation set, and 79.57% on the test set. The results show that the developed model can learn distinctive structural features in retinal OCT images and offers a feasible performance for multi-class classification. However, the difference between training and test performance indicates that the model's generalization capacity needs improvement.

Keywords: Retinal Diseases, Optical Coherence Tomography (OCT), CNN.

I. INTRODUCTION

Retinal diseases are significant ophthalmological problems that can affect visual function to varying degrees and, in some cases, lead to progressive vision loss [1]. Early diagnosis and accurate assessment of retinal structures are crucial for developing appropriate treatment plans and monitoring disease progression [2]. In this process, optical coherence tomography (OCT) is a widely used imaging method for clinical evaluation due to its ability to visualize the retina's cross-sectional structure at high resolution [3]. Distinguishing morphological changes caused by different retinal diseases in OCT images can be time-consuming, even for experienced specialists [4]. Furthermore, the increasing imaging volume has increased the need for computer-aided decision systems capable of automatically analyzing images. In recent years, the widespread use of deep learning-based methods in medical image analysis has enabled the direct learning of distinguishing features from OCT images.

Convolutional neural networks (CNNs) offer an effective approach to image classification problems by learning hierarchical representations of edges, textures, shapes, and more complex spatial patterns. While traditional methods require the predetermination of features and the design of a separate feature-extraction phase, CNN-based models can directly learn the necessary representations for image classification during training. This feature provides a significant advantage in evaluating retinal OCT images, where morphological differences can vary across classes. In this study, a specially designed CNN model was used to automatically classify retinal OCT images into 8 classes. The eight classes considered in the study were Age-related Macular Degeneration (AMD), Choroidal Neovascularization (CNV), Central Serous Retinopathy (CSR), Diabetic Macular Edema (DME), Diabetic Retinopathy (DR), DRUSEN, Macular Hole (MH), and NORMAL. The model's performance was evaluated using training, validation, and independent test subsets; the model's learning and generalization behavior was analyzed by examining changes in accuracy and loss during training.

II. LITERATURE REVIEW

The use of optical coherence tomography (OCT) images in the automated classification of retinal diseases has become a significant research area with the development of deep learning methods [5]. The ability of OCT images to reveal detailed structural changes in retinal layers makes this imaging method particularly prominent in the computer-aided diagnosis of macular diseases. A review of the literature reveals that research focuses not only on improving classification accuracy but also on issues such as noise tolerance, generalizability, explainability, and the integration of diverse data sources. Liew and Agaian, in a broad study of OCT-based ocular disease prediction methods, compared manual feature extraction methods used over the last two decades with deep learning-based approaches [6]. The study evaluated various disease groups, including glaucoma, diabetic retinopathy, cataracts, amblyopia, and macular degeneration; existing methods were examined for accuracy, sensitivity, specificity, and computational cost. The researchers also highlighted the importance of open-access OCT datasets in model development and discussed the security vulnerabilities of OCT-based diagnostic systems against adversarial attacks. In this respect, the study reveals that not only high accuracy but also model reliability and safety should be considered when classifying retinal disease. Redesigning network architectures to extract more distinctive features from OCT images is also widely discussed in the literature. Karthik and Mahadevappa proposed an approach that aims to make high-contrast regions, such as retinal layers, more prominent in feature maps by modifying residual connections in the standard ResNet architecture [7]. The developed structure increases the contrast in residual connections and enables the acquisition of sharper feature maps. In addition, the proposed activation function is designed to support smaller gradients while preserving negative weights. In experiments conducted on two different OCT datasets with four and eight classes, an increase in classification accuracy was reported compared to standard structures. In particular, an improvement of approximately 1.6% was observed when the initial accuracy was lower, and more significant gains were evident in the confusion matrix results when the class distribution was unbalanced, indicating that architectural improvements can be effective for classes with limited samples. Similarly, Peng, Lu, Zhuo, and Li focused on reducing the impact of speckle noise on classification performance in OCT images [8]. The researchers developed a model called the Multi-Scale Noise-Reducing Residual Convolutional Network (MS-DRCN). The proposed structure consists of Soft-Denoising Blocks to suppress speckle noise, Multi-Scale Context Blocks to capture contextual information at different scales, and Feature Fusion Blocks to combine high- and low-resolution features. The model achieved classification accuracy of 96.4% and 96.5% in the OCT2017 and OCT-C4 datasets, respectively. Furthermore, experiments adding different levels of Gaussian and speckle noise to the test data showed that the proposed method achieved 0.6% to 2.9% higher accuracy than ResNet. These findings demonstrate that noise-tolerant feature extraction is important for model performance in OCT classification systems. In addition to high classification success, the explainability of the models' decision processes is also a prominent issue in recent studies. Wang et al. investigated the use of explainable artificial intelligence methods for detecting age-related macular degeneration. The study considered various ophthalmic imaging modalities, including color fundus photography, OCT, ultra-wide-field fundus imaging, and fundus angiography [9]. The researchers aimed to optimize the deep learning model using insights from XAI methods and to make AMD diagnostic systems more reliable. The findings show that explainable artificial intelligence can contribute not only to making model decisions more understandable but also to improving model robustness. This approach is important for increasing confidence in AI-based decision support systems in clinical applications.

Multimodal approaches are being developed to evaluate retinal diseases, moving beyond models based solely on image data. The VisionTrack system proposed by Zedadra et al. has a hybrid structure that integrates multiple data sources [10]. The system uses a CNN for image feature extraction, a Graph Neural Network for modeling relationships among clinical risk factors, and a Large Language Model for processing patients' medical reports. Evaluated on the RetinalOCT and RFMID datasets, the system achieved an accuracy of 0.980, an F1 score of 0.979, a recall of 0.978, and a precision of 0.979 on the RetinalOCT dataset. Accuracy was reported as 0.989 on the RFMID dataset. The results of the study demonstrate that the combined evaluation of imaging, clinical

information, and patient-reported textual data can achieve strong classification performance in predicting retinal disease. Overall, the literature shows that CNN-based methods are effective for classifying retinal OCT images, but research is increasingly moving towards more complex, multi-component structures. Improvements in residual linkages, noise reduction mechanisms, multi-scale feature extraction, explainable artificial intelligence, and multimodal data integration are among the main approaches used to enhance performance and reliability. However, increased architectural complexity leads to higher computational costs and more model parameters. In this context, evaluating the performance of a relatively open CNN architecture that can directly distinguish eight retinal classes from OCT images is important for establishing a basic experimental framework for comparison with more complex methods.

III. MATERIAL AND METHODS

1. Dataset

Retinal OCT Image Classification - C8 dataset was used in this study [11]. The entire dataset consists of 24,000 retinal OCT images and includes eight classes: AMD, CNV, CSR, DME, DR, DRUSEN, MH, and NORMAL. Sample data is shown in Figure 1. In the original dataset structure, there are 3,000 images per class: 2,300 in the training folders, 350 in the validation folders, and 350 in the testing folders.

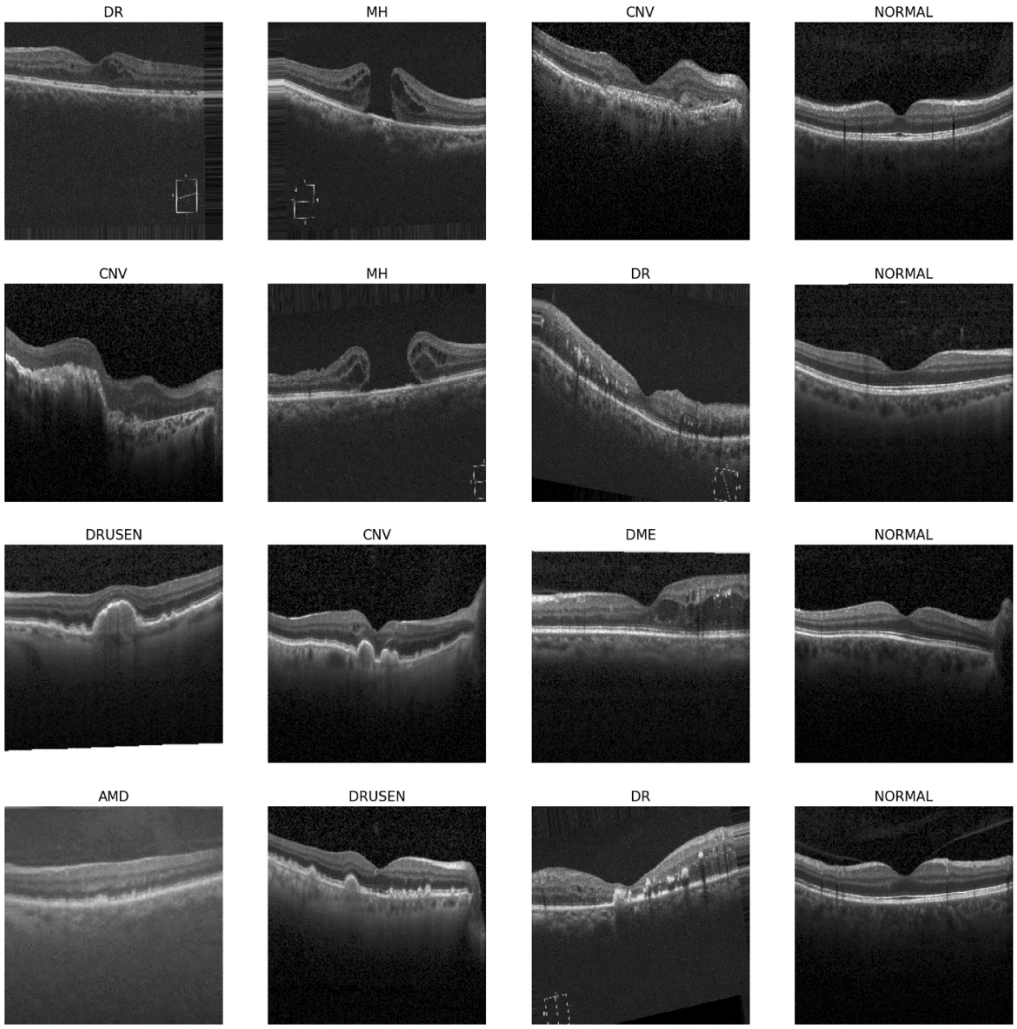


Figure 1. Sample data.

However, in this study, the dataset's pre-existing validation and test folders were not used directly. In the experimental design, only the 18,400 images found in the original train folder were used as the working data pool. Since each class contains 2,300 images, the dataset is perfectly balanced across classes. Stratified sampling was used to preserve the class distribution in dividing the data into training, validation, and test sets. In the first stage, 80% of the data was allocated to the training set, and the remaining 20% was split equally between the validation and test sets.

Table 1. Dataset ratios.

Dataset	**Image**	
Train	14.720	80%

Test	1.840	10%
Validation	1.840	10%
Total	18.400	100%

As a result of stratified division, each class was represented by 1,840 images in the training set, 230 in the validation set, and 230 in the test set. This ensured a balanced experimental structure, preventing the model performance from being affected by the numerical dominance of any single class. Since the original image sizes varied, all images were rescaled to 224 × 224 pixels and processed in three-channel RGB format before being fed into the model. The batch size was set to 16 for model training, and labels were created categorically, in accordance with the multi-class classification problem.

2. Convolutional Neural Networks

In this study, a custom convolutional neural network trained from scratch was developed to classify retinal OCT images. The model is constructed from three basic convolution blocks, each with increasing numbers of filters, to perform hierarchical feature extraction from low-level local features to more complex morphological structures in the images. The input layer of the model accepts RGB images with dimensions of 224 × 224 × 3. The first block contains two 3 × 3 Conv2D layers with 64 filters. The second block uses three Conv2D layers with 128 filters, and the third block uses four Conv2D layers with 256 filters. The same padding was used to preserve spatial dimensions across all convolution layers, and the Rectified Linear Unit (ReLU) activation function was used for nonlinear transformation. A 2 × 2 MaxPooling layer was applied after each convolution block. MaxPooling reduces the spatial dimensions of feature maps while preserving dominant features and reducing the computational load in subsequent layers. After the third convolution block, the resulting feature maps are converted into a one-dimensional feature vector via the Flatten layer. In the classification section, two fully connected layers with 256 and 64 neurons, respectively, are used, and ReLU activation is applied to each. The last layer consists of 8 neurons and uses the softmax activation function to produce class probability distributions. Based on the architecture defined in the code, the model has 53,870,600 trainable parameters. The general structure of the model architecture is shown in Figure 2.

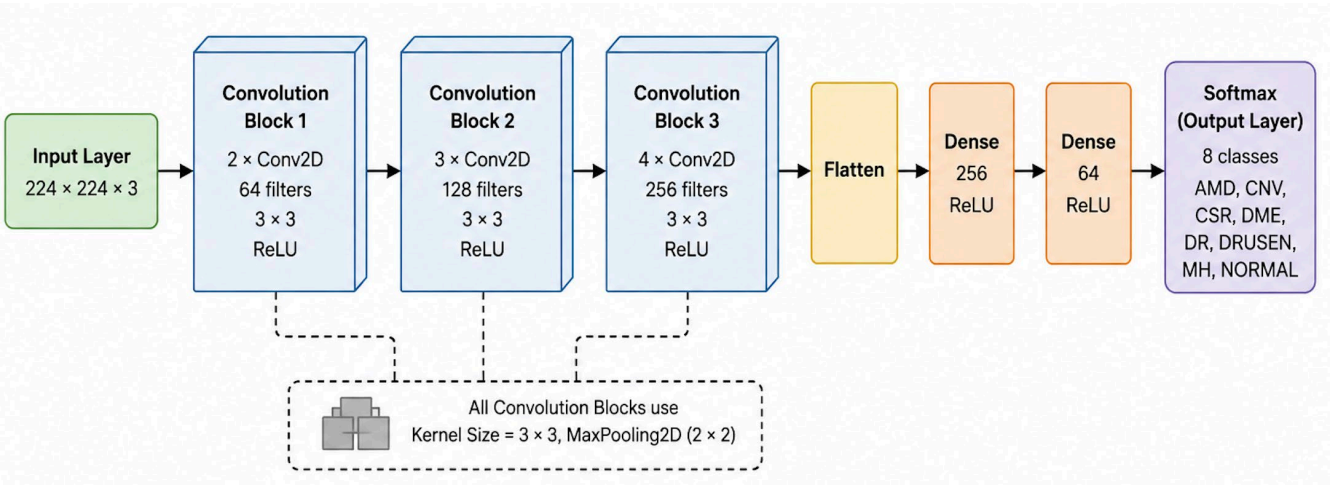


Figure 2. Model architecture.

IV. EXPERIMENT AND RESULTS

Experiments were conducted using the TensorFlow/Keras library. The training set used 14,720 images, while the validation and test sets each used 1,840 images. The batch size was set to 16, and the model was trained for 10 epochs. During training, at the end of each epoch, the accuracy and loss values on the validation set were calculated to monitor not only the model's learning on the training data but also its generalization to unseen data. As a result of training, the model achieved an accuracy of 0.9868 (98.68%) and a loss of 0.0380 on the training set. On the validation set, the accuracy was 0.8114 (81.14%) and the loss was 1.1539. On the test set, which was separate from the training process, the model showed an accuracy of 0.7957 (79.57%) and a test loss of 1.2704. The obtained values indicate that the model learned the patterns in the training data with a very high accuracy. However, the approximately 17.5 percentage-point difference between training and validation performance indicates that the model is becoming increasingly fit to the training data and that overlearning tendencies emerge as training progresses. The fact that test accuracy is close to validation accuracy is significant, as it shows that a similar level of generalization is achieved across both validation and test sets.

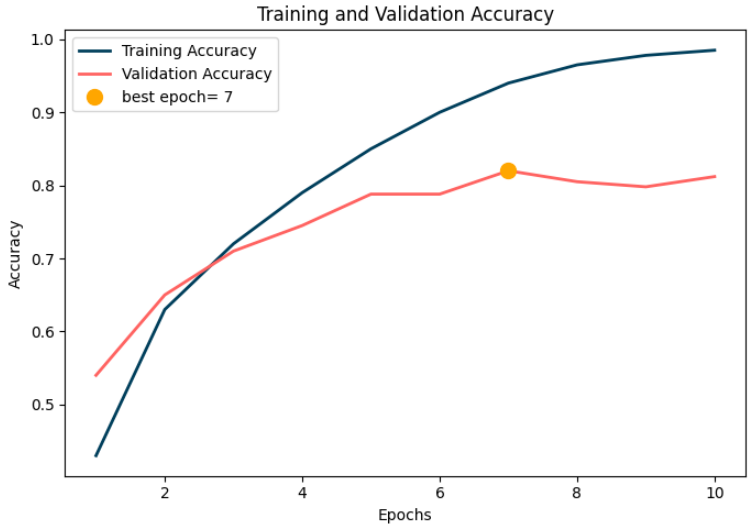


Figure 3. Accuracy curves.

As shown in Figure 3, training accuracy increased steadily from the first epoch and approached approximately 99% in the last epoch. Validation accuracy also increased significantly in the first part of training; however, the rate of increase decreased significantly after approximately the fifth epoch. The highest validation accuracy was achieved at approximately 82% in the 7th epoch. After this point, while training accuracy continued to increase, the fact that validation accuracy did not improve at the same rate indicates that training and validation performances began to diverge.

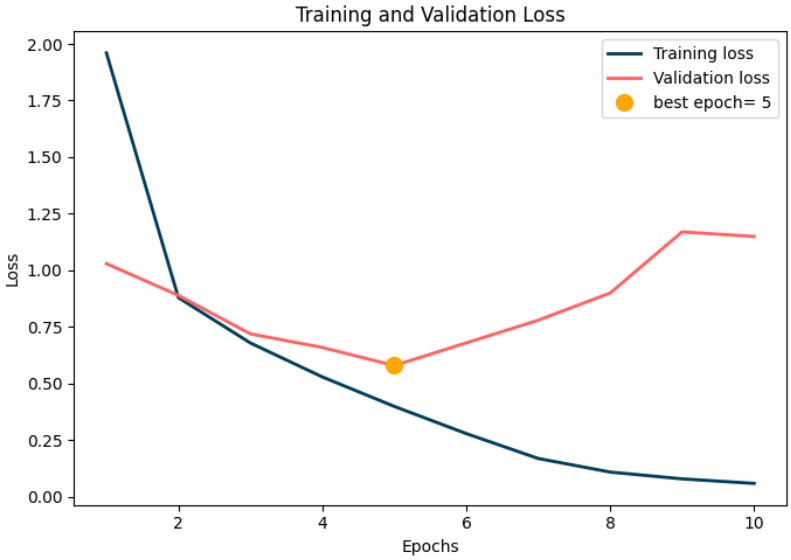


Figure 4. Loss curves.

A similar behavior is more clearly seen in the loss curves. Examining Figure 4, it can be seen that training loss steadily decreases throughout all epochs, approaching approximately zero. Validation loss also decreased during the first five epochs, reaching its lowest value around the 5th epoch. The fact that training loss continues to decrease while validation loss starts to rise again after the fifth epoch provides strong evidence that the model has begun overlearning. It is noteworthy that validation accuracy peaks around the 7th epoch, while validation loss reaches its minimum around the 5th epoch. While accuracy only considers the proportion of correctly classified samples, categorical cross-entropy also accounts for the model's predicted probabilities for correct and incorrect predictions. Therefore, even if the model continues to classify some images correctly, validation loss can increase as it begins producing higher-confidence values for incorrect predictions. Therefore, evaluating the training curves together is more informative than evaluating model behavior solely by accuracy. The overall test accuracy of 79.57% demonstrates that the developed CNN model can distinguish eight retinal OCT classes with reasonable generalization. However, despite training accuracy exceeding 98%, validation and test results remaining at around 80% indicate that the current architecture's high capacity leads to overfitting on the training data. The model's approximately 53.9 million trainable parameters are also consistent with this behavior; the fully coupled section, especially after the Flatten layer, constitutes a significant portion of the parameter count. These results suggest that methods such as early stopping, dropout, weight regularization, online data enhancement, or a lower-parameter classification header could be evaluated in future studies for stronger generalization performance. In particular, since validation loss begins to increase after the fifth epoch, saving the model weights at the best validation performance may yield more stable results than using the last-epoch model.

V. DISCUSSION

The findings show that the proposed CNN architecture achieves high training success on retinal OCT images, but validation and test performance decline significantly. Specifically, the fact that training loss continues to decrease while validation loss increases after the fifth epoch suggests that the model is overlearning. This may be related to the architecture's high capacity relative to the data volume, with approximately 53.9 million

trainable parameters. Nevertheless, the 79.57% test accuracy indicates that the model demonstrates strong classification performance on unseen OCT images. Future studies aim to improve generalization performance by increasing data volume, using dropout, early stopping, and employing fewer-parameter architectures.

ACKNOWLEDGEMENT

This study was conducted on a computer funded by the Gazi University Scientific Research Projects Coordination Unit under project number FPD-2025-9669.

REFERENCES

[1] Kaur, G., & Singh, N. K. (2022). Inflammation and retinal degenerative diseases. Neural regeneration research, 18(3), 513.

[2] MacGillivray, T. J., Trucco, E., Cameron, J. R., Dhillon, B., Houston, J. G., & Van Beek, E. J. R. (2014). Retinal imaging as a source of biomarkers for the diagnosis, characterization, and prognosis of chronic illnesses or long-term conditions. The British journal of radiology, 87(1040), 20130832.

[3] Hee, M. R., Izatt, J. A., Swanson, E. A., Huang, D., Schuman, J. S., Lin, C. P., ... & Fujimoto, J. G. (1995). Optical coherence tomography of the human retina. Archives of Ophthalmology, 113(3), 325-332.

[4] Muchuchuti, S., & Viriri, S. (2023). Retinal disease detection using deep learning techniques: a comprehensive review. Journal of Imaging, 9(4), 84.

[5] Alqudah, A. M. (2020). AOCT-NET: a convolutional network automated classification of multiclass retinal diseases using spectral-domain optical coherence tomography images. Medical & biological engineering & computing, 58(1), 41-53.

[6] Liew, A., & Agaian, S. (2025). Comprehensive Survey of OCT-Based Disorders Diagnosis: From Feature Extraction Methods to Robust Security Frameworks. Bioengineering, 12(9), 914.

[7] Karthik, K., & Mahadevappa, M. (2023). Convolution neural networks for optical coherence tomography (OCT) image classification. Biomedical Signal Processing and Control, 79, 104176.

[8] Peng, J., Lu, J., Zhuo, J., & Li, P. (2023). Multi-scale-denoising residual convolutional network for retinal disease classification using OCT. Sensors, 24(1), 150.

[9] Wang, M. H., Chong, K. K. L., Lin, Z., Yu, X., & Pan, Y. (2023). An explainable artificial intelligence-based robustness optimization approach for age-related macular degeneration detection based on medical IOT systems. Electronics, 12(12), 2697.

[10] Zedadra, A., Salah-Salah, M. Y., Zedadra, O., & Guerrieri, A. (2025). Multi-modal AI for multi-label retinal disease prediction using OCT and fundus images: a hybrid approach. Sensors, 25(14), 4492.

[11] Obuli Sai Naren. (2021). Retinal OCT Image Classification - C8 [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2736749

International Conference on Computer Sciences (IHCONCS 2026)
17-18 September 2026, Zagreb-Croatia

AI-assisted Error Analysis to Design a Writing Course Syllabus

Tuğba Dörterler
Graduate School of Educational Sciences
Gazi University *(English Language Teaching)*
Ankara, Türkiye
tdorterler@gazi.edu.tr
ORCID:0000-0003-3243-8181

Büşra Duygu Çelik
Graduate School of Natural and Applied Sciences,
Gazi University, *(Computer Engineering)*
Ankara, Türkiye
busraduygucelik@gazi.edu.tr
ORCID: 0000-0001-9640-1025

Abstract

This study aims to identify students' recurring writing errors through AI-assisted analysis and use the findings to design a writing course syllabus. To define course objectives, it is essential to explore students' needs. To design the syllabus for a writing course, 50 essays written by engineering students are analyzed by using Gen-AI technology (GPT-5.6), and the errors are identified under a pre-defined 18-category taxonomy. The findings indicate that the major language knowledge flaw lies in lexico-grammatical accuracy, particularly in collocation and subject-verb agreement, rather than in surface-level writing mechanics such as spelling or punctuation. The study recommends that the suggested course syllabus focus primarily on vocabulary, with an emphasis on teaching and learning vocabulary in chunks.

Keywords: syllabus design, error analysis, error frequency, Artificial Intelligence

I. INTRODUCTION

In academia, writing skills are considered a basic requirement for university students to convey their scientific understanding and participate in academic discussions in writing. Some faculties at Gazi University, including Medicine, Architecture, and Engineering, offer English-medium instruction. At that point, writing skills gain importance for those majors. The current writing course content is becoming obsolete; therefore, it needs to be more motivating and to the point in terms of students' needs. The students from the engineering department, despite being outstanding, need to improve their writing skills. The students receive English-medium instruction in their departments, so they mostly need to express their ideas in writing. Therefore, the new writing syllabus must meet their needs. This study aims to revise the writing course syllabus based on the results of AI-assisted error analysis of students' essays. The dynamic nature of technology requires some revisions to the current syllabus content. There are numerous types of AI-assisted applications for writing academic texts. However, students are required to produce original work related to their majors, so they need to improve their writing skills. To tailor a syllabus to students' needs, it is essential to conduct an error analysis of their own work.

Given the current issue, the following research questions are posed to achieve the study's aim.

- What types of linguistic errors are found in university EFL students' essays?
- Which error types are most prevalent in terms of raw and normalized frequencies?
- What do these error patterns suggest for the design of academic writing instruction?

1. Significance of the study

This study presents results indicating that instructors should maintain more learner-centered courses to improve students' writing skills.

II. RELATED LITERATURE

Recent developments in Gen-AI technologies infuse every aspect of education, including language learning and teaching. AI-assisted applications can be used in a wide range of areas in language education. In a meta-analysis by Xu et al. [1], it was found that AI-assisted second-language learning has a positive effect on receptive skills, namely listening and reading, as well as on learners' vocabulary knowledge. In addition to studies on the effects of AI-assisted language learning, numerous studies examine the benefits of AI technologies in second-language teaching, including course preparation [2,3], learner assessment [4,5], and language-learning material development [6].

1. Error Analysis in Sylabus Design for Language Teaching

Error analysis is a systematic approach to understanding errors made by second-language learners. Error analysis is conducted to understand what is missing in learners' language knowledge. It involves categorizing errors and determining causes, such as flaws in understanding grammatical rules. To design a learner-centered syllabus for language courses, it is essential to acknowledge learners' language levels.

2. AI-Assisted Error Analysis

AI technologies are primarily used to detect language errors and provide users with feedback to improve their writing skills, according to recent related literature. As in Tseng and Lin [7], AI-assisted writing tasks enhance students' writing skills and improve writing quality. Pack et al. [8] claim that with the development of AI technologies, providing feedback on written work will no longer be time-consuming in the near future. On the contrary, in this study, the main task for Gen-AI is the identification and classification of errors in the essays using a fixed, researcher-controlled taxonomy. After reviewing the related literature, there is no similar use of AI for this purpose.

III. METHODOLOGY

1. Participants and Writing Corpus

In this study, a corpus of learners' written samples was compiled from 50 essays (25 males and 25 females). Students study on a full-time basis at the Engineering Faculty from six different departments, including Industrial Engineering, Mechanical Engineering, Civil Engineering, Electrical and Electronics Engineering, Computer Engineering, and Chemical Engineering. The students have been studying at different departments of the English-medium Engineering Faculty for two years. In their first year, they have a three-hour General English course during two semesters. In their second year, they have a three-hour Academic English writing course during two semesters. Engineering students must take the Prep school proficiency exam before starting their majors. Students who pass the proficiency exam begin their own majors. Students who fail the proficiency exam must take a one-year preparatory language program with a 24-hour weekly schedule. The preparatory school curriculum offers A1, A2, and B1-level English language education based on a skills-based curriculum. The ultimate purpose is for students to succeed in B1-level language education across all four language skills equally.

Consequently, the students who take language courses at their own faculty have diverse language backgrounds. At that point, it is essential to design a syllabus shaped by students' needs.

2. Error Analysis Framework

Engineering students have written different types of essays, ranging from a comparison essay to an argumentative essay, during their essay quiz. The corpus is generated from fifty randomly selected essays across over three hundred quizzes. The essays are anonymized and photographed for upload to the AI system. To analyze the errors, the error categories are defined as shown in the table below.

Table 1. Error Categories

Code	Error Category	Operational Definition
LEX	Lexical choice	Incorrect/inappropriate lexical item for the intended meaning
COLL	Collocation/phraseology	Words are individually possible, but form an unconventional combination
WF	Word form	Correct lexical family but incorrect morphological form/category
ART	Article/determiner	Omission, unnecessary use, or incorrect choice of article/determiner
PREP	Preposition	Incorrect, missing, or unnecessary preposition
SVA	Subject-verb agreement	The verb does not agree with the grammatical subject
COUNT	Countability/number	Incorrect singular/plural or count/non-count treatment

Code	Error Category	Operational Definition
VF	Verb form/pattern	Incorrect tense, aspect, infinitive, gerund, participle, auxiliary, or complementation pattern
SYN	Sentence structure/syntax	Malformed clause or sentence structure not adequately explained by a more specific category.
PRO	Pronoun/reference	Incorrect, ambiguous, missing, or unclear pronoun/reference
ADVADJ	Adjective/adverb modification	Incorrect adjective/adverb form or inappropriate modification
CONJ	Conjunction/linking expression	Grammatically or semantically inappropriate connective
WO	Word order	Constituents appear in an unacceptable/non-target order
OM	Missing element	A necessary grammatical element is absent and does not belong more specifically to ART/PREP/etc.
RED	Redundancy	Unnecessary repetition or grammatically redundant element
PUNC	Punctuation	Missing, incorrect, or unnecessary punctuation
CAP	Capitalization	Incorrect upper/lower case

Code	Error Category	Operational Definition
SPELL	Spelling	An orthographic error that does not create a legitimate different lexical item

The images of the essays are named as M1, M2, etc. for male students, and F1, F2, etc. for female students. The images are uploaded in groups of five to the AI system. It works on the essays to identify errors and count them, enabling error frequency analysis across the 18 error categories. The system provides two datasets, including raw frequency and normalized frequency per 100 words for each essay.

3. AI-Assisted Analysis Procedure

Fifty anonymized essays are randomly selected from three hundred papers written by engineering students from different departments. Each paper has two different tasks, one of which is chosen by the student. Students have to write a five-paragraph essay on the given topic, with a word limit of 200-300 words. The essays are uploaded as images to ChatGPT one by one. They are first transcribed from handwritten to a digital form.

In the first step, the following prompt is provided in the ChatGPT chat box, along with images of five essays.

Could you please first transcribe the handwritten texts in the images? Then apply error analysis to each essay separately. At last, generate a table including error types and their raw and normalized frequencies for each one.

The system executes the prompt and first transcribes the essay in the image without correcting any mistakes. Afterward, it presents a table showing raw and normalized frequencies. The following image is a sample of the results.

F2 error analysis

Representative examples include *"Everyone need"* (SVA), *"theirself"* (PRO), *"In that point"* (PREP), *"first prefer"* (WF), *"an information"* (COUNT), *"so many information"* (COUNT), *"which one of the acceptable"* (SYN), *"the some signs"* (ART), *"reliable informations"* (COUNT), and *"help you for reality"* (COLL).

Code	Error type	Raw frequency	Per 100 words
SYN	Sentence structure / syntax	6	2.50
COUNT	Countability / number	6	2.50
ART	Article / determiner	5	2.08
SPELL	Spelling	5	2.08
SVA	Subject–verb agreement	3	1.25
PREP	Preposition	3	1.25
WF	Word form	2	0.83

With the help of the table, it is easy to identify which kinds of errors each student makes most and least. Therefore, it is possible to give individualized feedback to each student easily.

All five essay groups have been analyzed using the same prompt by the AI system. Afterward, the following prompt is used, and the system combines the tables into one spreadsheet in two versions. In the first version, all participants' errors were listed according to their pre-coded taxonomy. Below are the results from the 50-participant corpus.

Participant	Lexical Choice		Collocation		Word Form		Article		
	Raw	Norm/100	Raw	Norm/100	Raw	Norm/100	Raw	Norm/100	…
F1	3	1.13	8	3.01	2	0.75	3	1.13	
F2	1	0.42	2	0.83	2	0.83	5	2.08	
F3	0	0.00	6	1.93	0	0.00	5	1.61	
F4	1	0.49	6	2.94	6	2.94	6	2.94	
⋮									

All the data is given in an Excel document. After error analysis is applied to the fifty essays and a table is produced showing the number of errors and the normalized frequencies per 100 words, the following prompt is given to the AI system to generate more compact data identifying which error types are more prevalent in terms of normalized frequency.

Can you make a bar chart that sorts the error types from most to least prevalent using the Excel page uploaded here?

Gen-AI can read different kinds of documents and interpret the data in them. It is very efficient to convert the data from one format to another.

IV. RESULTS

1. Error Frequencies

With the assistance of AI technology, the number of mistakes categorized under the 18-category taxonomy is visualized into a bar chart in Figure 1, and the normalized frequency of the errors is demonstrated in Figure 2.

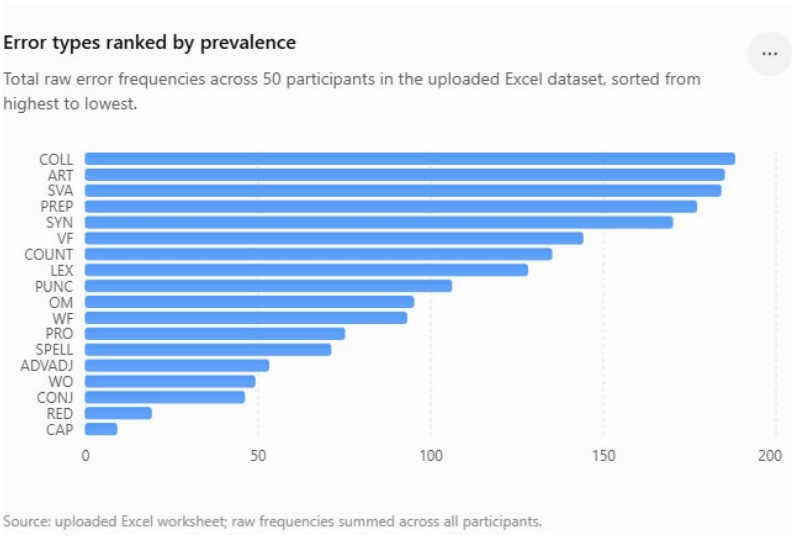


Figure 1. Raw error frequencies across the 18 error categories in the 50-participant corpus, ranked from highest to lowest.

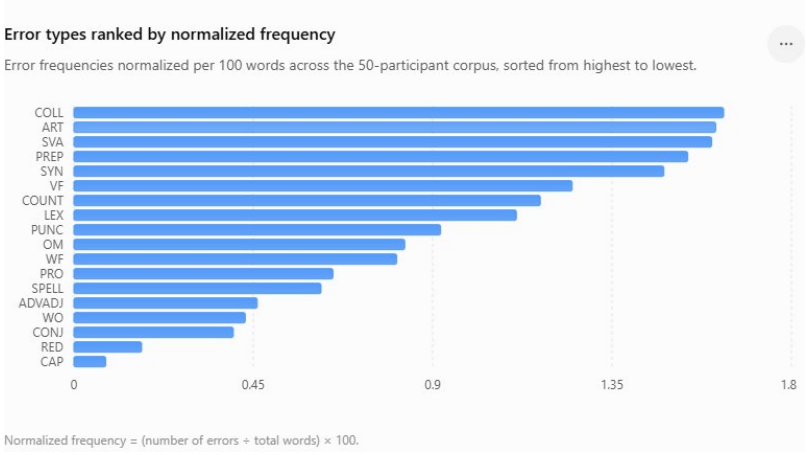


Figure 2. Normalized error frequencies (per 100 words) across the 18 error categories, ranked from highest to lowest.

As illustrated in Figures 1 and 2, the distribution of errors is not uniform across the 18 categories. Collocation errors were the most frequent (n = 188-1,63), closely followed by article/determiner (n = 185- 1,61), subject–verb agreement (n = 184-1,6), preposition (n = 177-1,54), and syntactic errors (n = 170-1,48). In contrast, redundancy (n = 19) and capitalization (n = 9) are relatively rare. The predominance of collocational, grammatical, and syntactic errors suggests that the learners' major difficulties lie in selecting contextually appropriate lexical combinations and controlling grammatical structures, rather than in surface-level writing conventions. The findings therefore indicate that EFL writing instruction may need to place greater emphasis on contextualized grammar and lexical patterning, particularly collocations, article use, agreement, and prepositions.

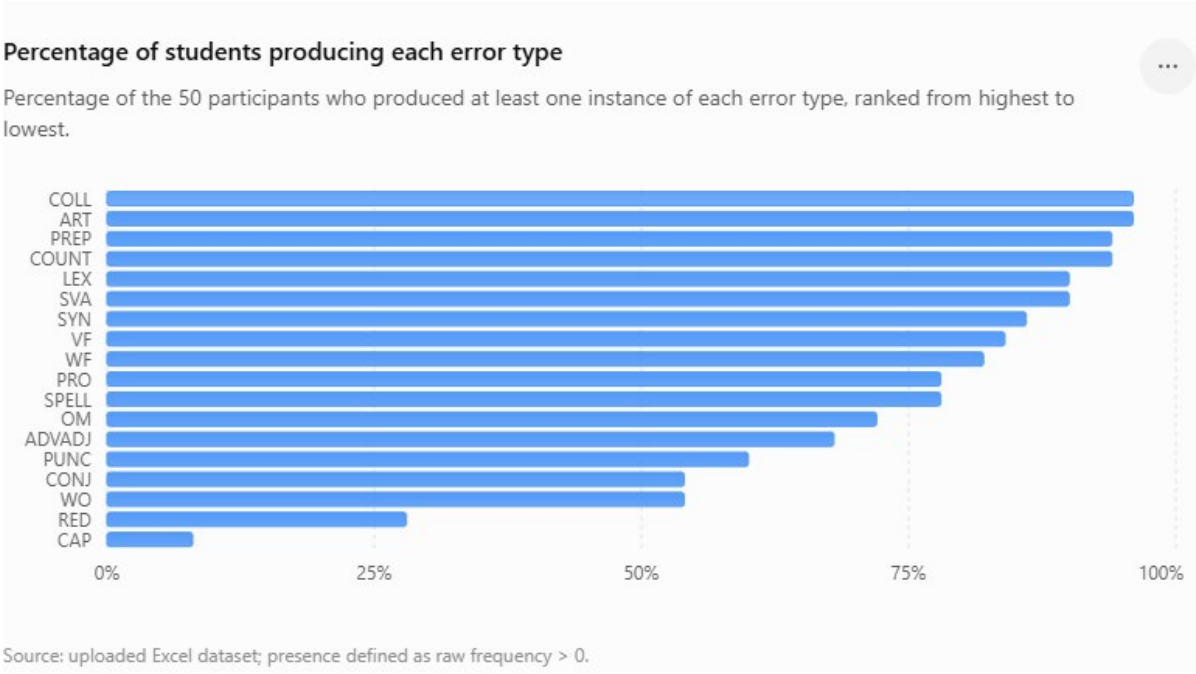


Figure 3. Percentage of the 50 participants producing at least one instance of each error type, ranked from highest to lowest.

Figure 3 shows that several error types were highly widespread across the 50 students. It indicates that they represent shared difficulties rather than problems restricted to a small number of learners. Collocation (COLL) and article (ART) errors are the most prevalent, occurring in 96% of the students, followed by preposition (PREP) and countability (COUNT) errors at 94**%**. Lexical (LEX) and subject–verb agreement (SVA) errors are also very common, each appearing in 90% of the participants. These findings suggest that difficulties involving both lexico-grammatical choice and grammatical accuracy are characteristic of the corpus as a whole.

A second group of errors is also broadly distributed. Syntax (SYN), verb form (VF), word form (WF), pronoun (PRO), and spelling (SPELL) errors occur in approximately 78–86% of the students. Thus, although their total raw frequencies may differ, these problems are experienced by a substantial majority of the learners.

In contrast, some error types are considerably less widespread. Punctuation (PUNC) occurs in 60%, while conjunction (CONJ) and word-order (WO) errors each occur in 54%. The least prevalent categories are redundancy (RED), found in 28% of students, and capitalization (CAP), found in only 8%.

V. DISCUSSION

The strongest learner need appears to be lexico-grammatical accuracy rather than writing mechanics alone**.** In particular, COLL, ART, PREP, and COUNT deserve the highest instructional priority because they combine high raw error frequencies with very high prevalence across students**.** This convergence is important: the errors are not simply repeated many times by a small group of weaker writers; almost the entire learner population shares them.

A second major need concerns sentence-level grammatical control, particularly subject–verb agreement, syntax, verb forms, and word formation. These results suggest that learners possess sufficient linguistic resources to produce extended written texts, yet still experience difficulty accurately combining them at the phrase and sentence levels.

Consequently, findings point toward three broader learner needs: **(**1) stronger phraseological and lexical competence, especially collocations; (2) greater control of grammatical form, particularly articles, countability, prepositions, agreement, and verb forms; and (3) improved self-editing and proofreading skills for spelling, omissions, punctuation, and other lower-frequency errors.

VI. CONCLUSION

With Gen-AI, large datasets can be easily and efficiently managed to provide a broad picture of the current situation across every aspect of education, from material development to the evaluation of learners' output. With this technology, time-consuming tasks are handled more efficiently in terms of time and human resources. Although some ethical issues related to Gen-AI exist, such as academic dishonesty [9] and data privacy [10], it facilitates big data management. It motivates teachers to improve the quality of their instruction. It helps reduce cognitive workload, enabling teachers to develop more creative ways to implement their courses.

REFERENCES

[1]. G. Xu, A. Yu, and L. Liu, "A meta-analysis examining AI-assisted L2 learning," *International Review of Applied Linguistics in Language Teaching*, no. 2, pp. 905-937, 2025, doi: 10.1515/iral-2024-0213.

[2]. X. Guo and H. Chen, "Incorporating generative AI in foreign language teaching preparation: An empirical investigation into the effects of AI-assisted lesson planning," *International Journal of Advanced Computer Science and Applications*, vol. 17, no. 4, Jan. 2026, doi: 10.14569/IJACSA.2026.0170428.

[3]. M. Zaim et al., "AI-powered EFL pedagogy: Integrating generative AI into university teaching preparation through UTAUT and activity theory," *Comput. Educ. Artif. Intell*., vol. 7, art. no. 100335, 2024, doi: 10.1016/j.caeai.2024.100335.

[4]. M. Y. M. Amin, "AI and Chat GPT in language teaching: Enhancing EFL classroom support and transforming assessment techniques," *International Journal of Higher Education Pedagogies*, vol. 4, no. 4, pp. 1-15, Dec. 2023, doi: 10.33422/ijhep.v4i4.554.

[5]. J. Kic-Drgas and F. Kılıçkaya, "Using artificial intelligence (AI) to create language exam questions: A case study," *XLinguae*, vol. 17, no. 1, pp. 20-33, 2024, doi: 10.18355/XL.2024.17.01.02.

[6]. D. B. Kır and E. Tekin, "Key features of AI applications in foreign language teaching," *BUEFAD*, vol. 14, no. 4, pp. 1143-1162, Oct. 2025,doi: 10.14686/buefad. 1681050.

[7]. Y.C. Tseng and Y.H. Lin, "Enhancing English as a Foreign Language (EFL) learners' writing with ChatGPT: A university-level course design," *Electronic Journal of e-Learning*, vol. 22, no. 2, pp. 78-97, 2024, doi: 10.34190/ejel. 21.5.3329.

[8]. A. Pack, K. J. Hartshorn, J. Escalante, and N. Gillette, "How well can GenAI (GPT-4) provide written corrective feedback on English-language learners' writing?" *International Journal of English for Academic Purposes: Research and Practice*, vol. 5, no. 1, pp. 7-26, 2025, doi: 10.3828/ijeap. 2025.2.

[9]. E. Mohammadkarimi, "Teachers' reflections on academic dishonesty in EFL students' writings in the era of artificial intelligence," *Journal of Applied Learning & Teaching*, vol. 6, no. 2, pp. 105–113, 2023, doi: 10.37074/jalt. 2023.6.2.10.

[10]. L. Huang, "Ethics of artificial intelligence in education: Student privacy and data protection," *Sci Insights Educ Front*, vol. 16, no. 2, pp. 2577–2587, Jun. 2023, doi: 10.15354/sief. 23. re202.