Dermatology Practical & Conceptual

An Explainable Hybrid Deep Learning Framework Combining Fine-Tuned Convolutional and Transformer-Based Ensembles for Dermoscopic Images of Skin Lesions

Büşra Duygu Çelik[1,3], Burcu Bekac[2], Esra Adışen[2], Adem Tekerek[3]

1 Gazi University, Graduate School of Natural and Applied Sciences, Computer Engineering Department, Ankara, Turkey
2 Gazi University Faculty of Medicine, Dept. of Dermatology, Ankara, Turkey
3 Gazi University, Faculty of Technology, Dept. of Computer Engineering, Ankara, Turkey

Key words: Dermoscopy, Deep learning, Explainable artificial intelligence, CNN, XAI
Citation: Duygu Çelik B, Burcu Bekac B, Esra Adışen E, Adem Tekerek A. An Explainable Hybrid Deep Learning Framework Combining Fine-Tuned Convolutional and Transformer-Based Ensembles for Dermoscopic Images of Skin Lesions. *Dermatol Pract Concept*. 2026;16(S1):7739. DOI: https://doi.org/10.5826/dpc.16S1a7739
Accepted: September 3, 2026; **Published:** November 2026

Copyright: ©2026 Duygu Çelik et al. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (BY-NC-4.0), https://creativecommons.org/licenses/by-nc/4.0/, which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original authors and source are credited.

Corresponding Author: Büşra Duygu Çelik, PhD Candidate, Graduate School of Natural and Applied Sciences, Gazi University, 06560 Yenimahalle, Ankara, Türkiye. E-mail: busraduygucelik@gazi.edu.tr , ORCID ID: 0000-0001-9640-1025

ABSTRACT

Introduction: Accurate diagnosis in dermoscopic examination is challenging due to the visual similarity between skin lesion types and the variability within the same group. Early diagnosis and treatment of skin diseases, especially skin cancer, are vital.

Objectives: This study aimed to develop and evaluate an explainable multi-architecture framework integrating convolutional, transformer-based, hybrid, and ensemble learning approaches for multi-class skin lesion classification.

Methods: Using HAM10000, four architectures (EfficientNet-B0, MobileNetV3-Large, ResNet34, DeiT-Small) were fine-tuned in two phases and combined into a validation-optimized, macro-F1-driven soft-voting ensemble; an independent ResNet34+XGBoost hybrid was evaluated in parallel. Lesion-grouped partitioning was used to prevent overlap between data subsets, with an additional image-level experiment assessing the influence of partitioning strategy on performance. Evaluation used accuracy, macro-averaged precision, recall, F1-score, per-class metrics, nested 5-fold cross-validation, McNemar's test, and lesion-clustered bootstrap effect sizes (95% CIs). Grad-CAM, Grad-CAM++, and LIME provided interpretability, with Grad-CAM localization evaluated against segmentation masks via Intersection over Union (IoU).

Results: Under lesion-grouped evaluation, the ensemble achieved the highest performance, with 85.76% accuracy, 75.24% macro-F1, and 0.971 macro-AUC. Paired accuracy gains over comparators ranged from 2.36 to 7.57 percentage points, with all lesion-clustered 95% CIs excluding zero. ResNet34 yielded the highest mean Grad-CAM IoU (0.456).

Conclusion: The ensemble achieved improved classification performance under lesion-grouped evaluation. Quantitative localization analysis showed measurable spatial overlap with lesion regions, although clinical usefulness requires reader studies and external validation.

Introduction

Skin lesions are abnormal structures that exhibit morphological and structural differences from healthy tissue. While many skin lesions are benign, some are malignant and can pose serious health risks. Melanoma, in particular, one of the most dangerous types of skin cancer, causes high mortality rates when metastases develop [1]. However, when detected early, it is largely treatable, and mortality rates can be reduced. Therefore, accurate and timely diagnosis is vital.

In recent years, dermoscopy has significantly improved the visualization of subsurface structures and the evaluation of skin lesions. Parallel to these developments, deep learning (DL)-based methods are increasingly being applied to skin lesion classification tasks. These methods enable higher accuracy and more objective evaluation in clinical decision-making processes, reduce human error, and improve the likelihood of early diagnosis [2]. Accordingly, the development of DL-based systems capable of classifying both benign and malignant skin lesions has become an important research area. Despite their impressive performance, most DL models rely on black-box systems that generate predictions without providing insight into the underlying decision-making process. This lack of transparency is the most significant challenge to interpretability and clinical applicability. Therefore, to ensure clinical usability and user acceptability, it is crucial to explain the reasoning behind DL models' conclusions when they misclassify cases or yield unexpected results. To overcome this challenge, explainable artificial intelligence (XAI) techniques have been developed to provide visual and interpretable representations of model behavior [3]. Among these, gradient-based methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) are widely used to highlight the domains of interest that contribute to classification decisions [4].

This study aims to develop and evaluate an explainable multi-architecture framework for multi-class skin lesion classification. EfficientNet-B0, MobileNetV3-Large, ResNet34, and DeiT-Small were fine-tuned and combined using a validation-optimized soft-voting ensemble, while a ResNet34+XGBoost hybrid was evaluated separately. Lesion-grouped partitioning prevented same-lesion overlap across data subsets. Grad-CAM, Grad-CAM++, and LIME were used for model interpretation, with Grad-CAM localization quantitatively evaluated against segmentation masks using intersection over union (IoU).

Objectives

This study aimed to develop and evaluate an explainable multi-architecture deep learning framework for multi-class skin lesion classification by integrating convolutional and transformer-based models within an ensemble framework. Furthermore, Grad-CAM attributions were quantitatively evaluated against ground-truth segmentation masks, complementing the visual explanations provided by Grad-CAM++ and LIME.

Methods

Dataset Description

The open-access HAM10000 dataset was used to classify skin lesions in this study. The dataset, consisting of a total of 10,015 dermoscopic images, comprises seven classes representing different skin lesion types [dermatofibroma (DF), actinic keratosis-intraepithelial carcinoma (AKIEC), melanoma (MEL), basal cell carcinoma (BCC), benign nevus (NV), benign keratosis-like lesions (BKL), and vascular lesions (VASC)] [5]. Representative images of these classifications are given in Figure S1. Because HAM10000 may contain multiple images for the same lesion, dataset partitioning was performed using lesion identifiers rather than individual images to prevent overlap of the same lesion across training, validation, and test subsets.

Data Preprocessing and Augmentation

The dataset was partitioned into training, validation, and test subsets using lesion identifiers (lesion_id) to prevent same-lesion overlap across subsets, yielding approximately 66%/17%/17% of images (train/validation/test) as lesion grouping precludes exact target ratios; zero lesion-level overlap between subsets was verified. Class balancing and augmentation were restricted to the training set. Minority classes were oversampled to 6,000 samples per class, while the majority class was not undersampled. Images were resized to 224×224 pixels and augmented using the Albumentations library, following the strategy of Akyel and Arıcı, to account for geometric and illumination variability [6]. Spatial augmentations included horizontal and vertical flipping, random 90-degree rotation, and shift-scale-rotate transformations (shift limit 0.05, scale limit 0.05, rotation limit ±15°). Color-based augmentations included RGB channel shifting and random brightness-contrast adjustment (Table S1). All images were then normalized using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) and converted to tensors. Validation and test images were resized and normalized only, without augmentation or oversampling.

Deep Ensemble and Hybrid Model Development

The proposed framework integrates four base architectures for multi-class skin lesion classification: EfficientNet-B0, MobileNetV3-Large, ResNet34, and DeiT-Small, all initialized with ImageNet-pretrained weights. This accelerates the

learning of low-level visual features such as edge structures, texture, color gradients, and lesion boundaries.

EfficientNet-B0 was chosen for its compound scaling approach, which scales network width, depth, and resolution in a balanced manner; MobileNetV3-Large was preferred for its lightweight architecture and computational efficiency; and the ResNet34 architecture was included in the study because it mitigates the gradient propagation problem through residual connections and provides training stability. DeiT-Small, a data-efficient vision transformer that complements these convolutional architectures, was incorporated to capture global contextual relationships through self-attention, thereby increasing architectural diversity and enabling complementary feature representation within the ensemble [7,8].

Each architecture was fine-tuned separately using a two-phase strategy. In Phase 1, the backbone was frozen and only the final block and classification head were trained using the Adam optimizer (learning rate 1e-4) with early stopping based on validation loss. In Phase 2, the full backbone was unfrozen and fine-tuned at a lower learning rate (1e-5), continuing from the Phase 1 checkpoint with independent early stopping; the best checkpoint across both phases was retained. Softmax probabilities from the four fine-tuned models were combined through weighted soft voting. The ensemble weights (ResNet34 = 0.20, EfficientNet-B0 = 0.25, MobileNetV3-Large = 0.15, DeiT-Small = 0.40) were selected by grid search (step = 0.05) on the validation set, optimizing macro-averaged F1-score rather than accuracy. Five-fold lesion-grouped cross-validation was additionally performed for each architecture and for the ensemble; within each outer fold, ensemble weights were re-selected independently from that fold's own inner-validation partition and never from its held-out data, yielding a nested cross-validation estimate. Grad-CAM and Grad-CAM++ activation heatmaps were generated from the final convolutional layers of the three convolutional architectures (Grad-CAM is not applicable to the DeiT-Small transformer), and LIME visualized the superpixel regions that supported or contradicted each prediction (Figure 1).

In addition to Branch A, a separate hybrid branch (Branch B) was also evaluated. In this branch, 512-dimensional Global Average Pooling (GAP) features extracted from an independently fine-tuned ResNet34 encoder were used to train an XGBoost classifier. Unlike the ResNet34 model in Branch A, which was optimized for end-to-end classification, this encoder was trained solely for feature extraction, allowing a direct comparison between end-to-end deep learning and a deep-feature-plus-gradient-boosting approach. Unlike Branch A, Branch B was developed using the original, non-augmented training split without class balancing, thereby isolating the effect of the feature-extraction-plus-boosting strategy from the class-balancing and augmentation procedures applied to Branch A. The XGBoost classifier used 300 estimators, a maximum tree depth of 6, a learning rate of 0.05, and subsample and column-subsample ratios of 0.8. To clearly separate model development from performance estimation, the final hybrid model used for independent test-set evaluation and the fold-wise cross-validation procedure used to assess generalization were implemented as two distinct workflows. For the final hybrid model, a separately fine-tuned ResNet34 encoder was trained on the full original training split, after which GAP features were extracted and used to train a final XGBoost classifier, which was evaluated on the independent test set. The fold-wise evaluation employed a stratified, lesion-grouped cross-validation protocol, described in detail in the Statistical Analysis section below. The pseudocode describing both branches is presented in Table 1.

Statistical Analysis

Model performance was evaluated using multiple classification metrics, including accuracy, precision, recall, F1-score, Area Under the Curve(AUC), and receiver operating characteristic (ROC) curve analysis [9]. Confusion matrices were generated for each model to examine class-level prediction patterns, where true positives and true negatives denote dermoscopic images correctly assigned to, or correctly excluded from, a given lesion class, and false positives/negatives denote misclassified samples:

$$N = TP + FP + TN + FN \quad (1)$$

$$\text{Accuracy} = (TP + TN)/N \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

$$\text{Recall} = TP/(TP + FN) \quad (4)$$

$$F1_{Score} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (5)$$

All metrics were computed per class and then macro-averaged (unweighted mean across the seven diagnostic classes); macro-averaged F1-score is reported as the principal class-balanced summary metric, with per-class values reported separately (Table 3). Average precision was additionally computed per class to complement AUC under class imbalance.

Generalization was assessed by 5-fold lesion-grouped cross-validation for both branches, as described in the Methods. Lesion-level grouping (StratifiedGroupKFold for outer folds, GroupShuffleSplit for internal validation splits) was applied throughout, ensuring that neither the encoders nor the classifiers used to evaluate a given fold were influenced by that fold's held-out samples; automated checks

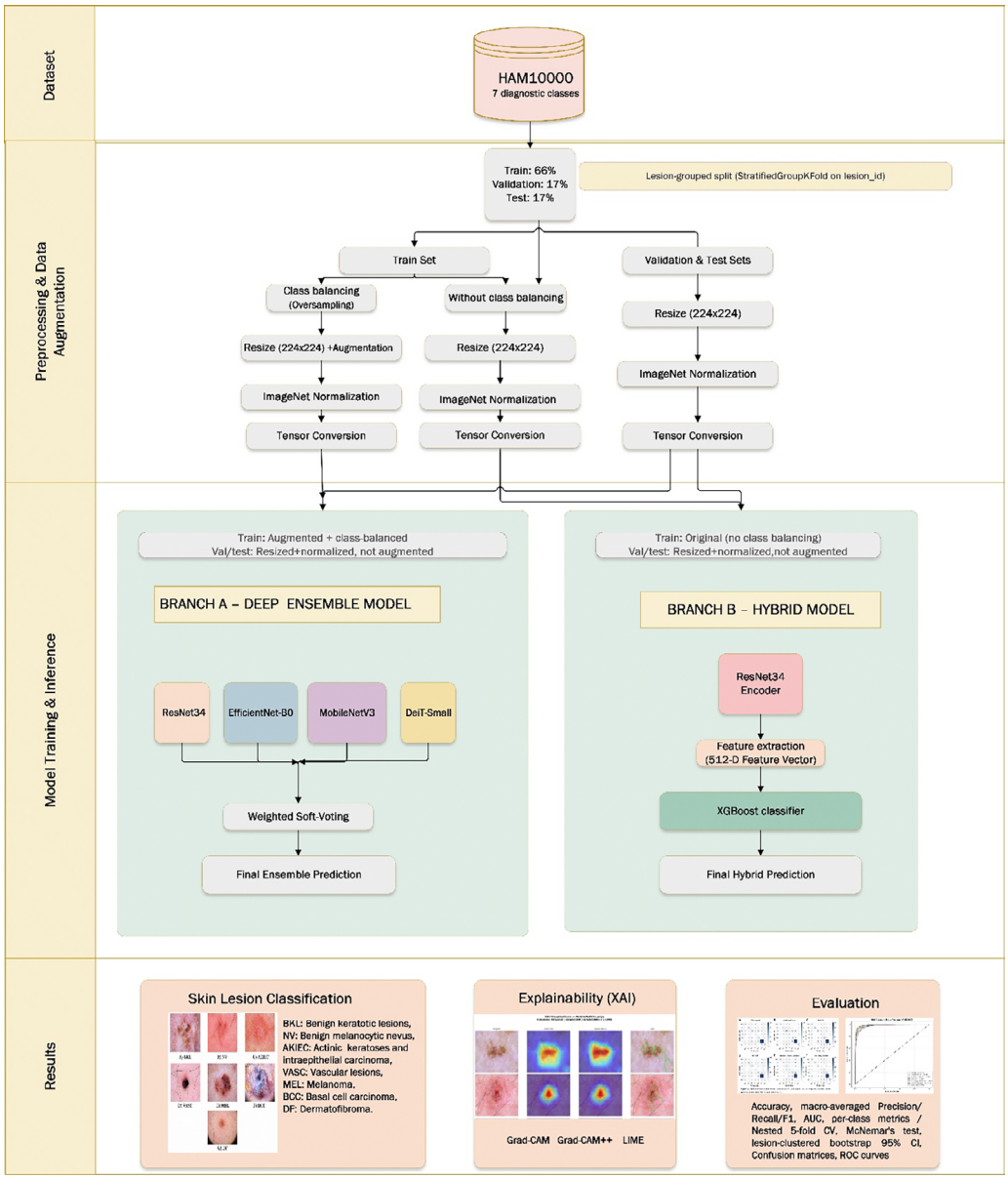

Figure 1. Schematic of the end-to-end classification framework, detailing the integration of XAI modules and ensemble voting within the experimental steps.

confirmed zero image- and lesion-level overlap across all split boundaries.

Differences between the ensemble and each comparator were assessed using McNemar's test on paired image-level predictions [10]. Because statistical significance alone does not convey the magnitude of improvement, paired accuracy differences were additionally reported with 95% confidence intervals obtained by lesion-clustered bootstrap resampling (10,000 resamples, resampling unique lesions with replacement to account for within-lesion correlation).

The spatial validity of Grad-CAM activations was evaluated quantitatively by computing the Intersection-over-Union (IoU) between binarized activation maps (targeting the ground-truth class) and the expert-annotated HAM10000 segmentation masks, for the three convolutional architectures [5,11].

Results

During the training and evaluation phase of the model, the Google Colab environment was used, utilizing the NVIDIA A100 GPU. Experiments were conducted using Python 3.12.12 and the IPython interactive shell.

Overall Classification Performance

Six approaches were evaluated on the independent HAM10000 test set: the four individual architectures, the proposed weighted soft-voting ensemble, and the hybrid model (Figure 2). Overall performance is summarised in Table 2; class-wise performance of the proposed ensemble is presented in Table 3, with corresponding values for the remaining models provided in Table S2.

Table 1. Algorithm of Hybrid Deep Ensemble with XAI .

Input*: D // HAM10000 Dataset of dermoscopic images*	
Output*: P_ens, P_hyb // Final diagnostic predictions and* M_{XAI} *// Interpretability heatmaps // IoU_score // Quantitative localization score* **Start**	
	1. ***Data Engineering & Preparation*** *1 Split D into training, validation, and testing sets using lesion-grouped stratified partitioning (~66%/17%/17% of images), ensuring all images of a lesion remain in a single subset. The original training split is retained unaltered as D_train for the hybrid branch (Step 3).* *2 Create a balanced training set D_train_bal by oversampling minority classes in D_train to 6,000 images per class; D_train itself remains unmodified.* *3 Resize, normalize, and augment D_train_bal (spatial and color transformations). D_train, validation, and test sets are resized and normalized only, without augmentation.*
	2. ***Branch A-Deep Ensemble*** *Initialize ResNet34, EfficientNet-B0, MobileNetV3-Large, and DeiT-Small with ImageNet-pretrained weights.* *1* ***for each*** *model* ***in*** *{ResNet34, EfficientNet-B0, MobileNetV3-Large, DeiT-Small}* ***do*** • *Phase 1: Freeze backbone; train final block and classification head with Adam (LR = 1e-4), early stopping on validation loss.* • *Phase 2: Unfreeze full backbone; continue fine-tuning with Adam (LR = 1e-5), independent early stopping.* • *Retain the best checkpoint across both phases.* *2 Determine ensemble weights (w_res, w_eff, w_mob, w_deit) via grid search (step = 0.05) on the validation set, maximizing macro-averaged F1-score.* *3 Validate generalization of each architecture and the ensemble via 5-fold lesion-grouped cross-validation; within each outer fold, re-select ensemble weights from that fold's inner-validation partition only (nested protocol).* *4 For each test image: compute softmax probabilities {p_res, p_eff, p_mob, p_deit}; combine via weighted soft voting:* *P_ens = w_res*p_res + w_eff*p_eff + w_mob*p_mob + w_deit*p_deit*
	3. ***Branch B- Hybrid model (independent of Branch A)*** *1 For each of five stratified, lesion-grouped folds drawn from D_train, fine-tune a fresh ResNet34 encoder (ImageNet-pretrained) using only that fold's training partition, with early stopping via an internal lesion-grouped validation split; extract fold-specific 512-dimensional GAP features; train an XGBoost classifier on the extracted features; and evaluate performance on the held-out fold.* *2 Report fold-wise cross-validation performance as the mean ± standard deviation across the five folds.* *3 Fine-tune a final ImageNet-pretrained ResNet34 encoder on the complete D_train; extract GAP features; train the final XGBoost classifier; and evaluate the hybrid model on the independent test set.* *4 Output hybrid prediction P_hyb.*
	4. ***Explainability (XAI) Generation.*** *1 for each predictedClass do* • *Generate Grad-CAM, Grad-CAM++ heatmaps (*M_{XAI}*) and LIME super-pixel explanations for each prediction.* • *Compute IoU between binarised Grad-CAM maps and HAM10000 segmentation masks for the three CNN architectures.*
	5. ***Evaluation & Performance Benchmarking.*** • *Calculate macro-averaged metrics: Accuracy, Recall, Precision, F1-Score, and AUC.* • *Generate confusion matrices and ROC curves.* • *Compare models using McNemar's test together with paired accuracy differences and lesion-clustered bootstrap 95% confidence intervals.*
	6. ***Return: P_ens, P_hyb, IoU_score,*** M_{XAI} ***and Performance Metrics.*** **End** //Area Under the Curve (AUC)//Receiver Operating Characteristic (ROC)//

The ensemble achieved the highest accuracy (85.76%) and macro-AUC (0.971), followed in accuracy by the hybrid model (83.39%), DeiT-Small (83.27%), EfficientNet-B0 (81.44%), ResNet34 (81.32%), and MobileNetV3-Large (78.19%). It also achieved the highest macro-precision (0.761), macro-recall (0.751), and macro-F1 (0.752), indicating more balanced performance across classes (Table 2).

Class-wise results showed a consistent pattern. The ensemble achieved the highest F1-score across all seven diagnostic categories compared with the individual architectures and the hybrid model (Table 3; Table S2), indicating that the

Table 2. Detailed performance results of all models.

Model	Accuracy	Macro-Precision	Mcro-Recall	Macro-F1-Score	Macro-AUC
ResNet34	81.32%	70.57%	67.74%	68.73%	95.48%
EfficientNet-B0	81.44%	67.48%	67.73%	67.11%	96.03%
MobileNetV3-Large	78.19%	66.41%	61.83%	62.78%	94.18%
DeiT-Small	83.27%	68.41%	73.00%	70.44%	95.69%
ResNet34 + XGBoost	83.39%	71.49%	62.59%	65.11%	94.88%
Ensemble (proposed)	85.76%	76.05%	75.08%	75.24%	97.12%

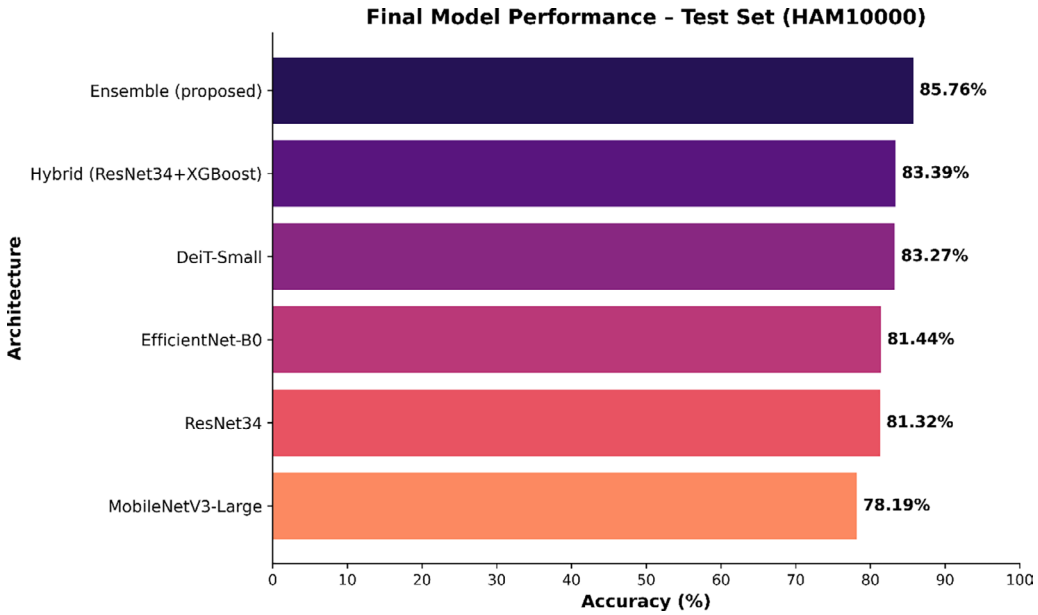

Figure 2. Comparative performance analysis highlighting the superior accuracy of the proposed soft-voting ensemble over standalone and hybrid architectures.

Table 3. Class-wise performance of the proposed weighted soft-voting ensemble on the independent test set.

Class	Precision	Sensitivity	F1-Score	AUC	n
AKIEC	0.589	0.541	0.564	0.967	61
BCC	0.681	0.861	0.761	0.989	72
BKL	0.753	0.665	0.706	0.953	197
DF	0.818	0.667	0.735	0.994	27
MEL	0.638	0.676	0.656	0.931	185
NV	0.939	0.942	0.940	0.968	1129
VASC	0.905	0.905	0.905	0.997	21

Abbreviations: BKL: Benign keratosis-like lesions, NV: Benign melanocytic nevus, AKIEC: Actinic keratosis-intraepithelial carcinoma, VASC: Vascular lesions, MEL: Melanoma, BCC: Basal cell carcinoma, DF: Dermatofibroma. n denotes the number of test images per class. Corresponding values for the individual architectures and the hybrid model are provided in Table S2.

performance gain was not confined to the majority class. Performance nevertheless varied substantially across categories: F1-scores ranged from 0.94 for NV (n=1129) to 0.56 for AKIEC (n=61), while melanoma achieved an F1-score of 0.66 with a sensitivity of 0.68. This variation explains the difference between overall accuracy and macro-F1 and indicates that minority-class discrimination remains a key limitation. A dissociation between ranking and decision performance was also observed: DF achieved a high AUC (0.994) despite a sensitivity of 0.667. This pattern was more pronounced for the hybrid model, whose DF sensitivity was 0.185 (F1=0.286) despite an AUC of 0.920, illustrating that strong probability ranking does not necessarily translate into correct classification at the operating threshold for rare classes.

ROC analysis (Figure 3) showed high threshold-independent discrimination across all six approaches, with macro-AUC values ranging from 0.942 for MobileNetV3-Large to 0.971 for the ensemble.

Confusion Matrix Analysis

Confusion matrices (Figure 4 A–F) showed that NV was classified most reliably by the ensemble (F1 = 0.94), whereas AKIEC (0.56), MEL (0.66), and BKL (0.71) remained more challenging. Performance was not directly associated with

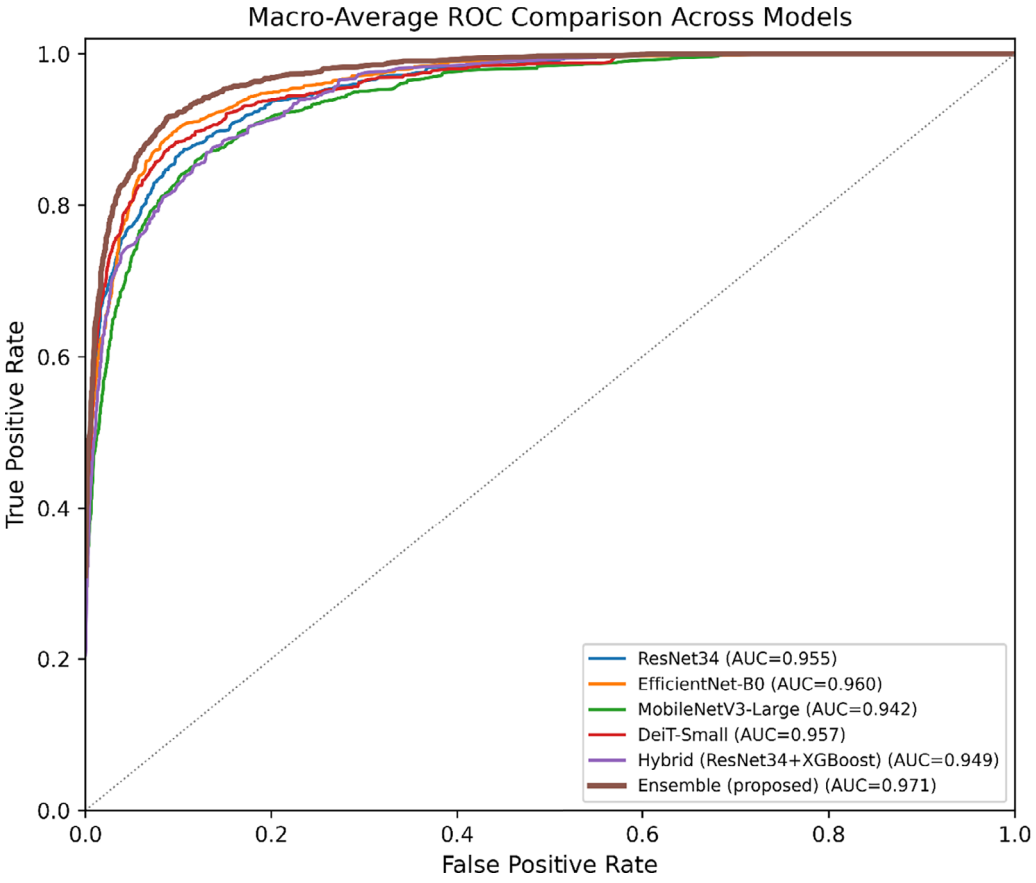

Figure 3. Macro-average ROC curves for all six evaluated classification models. The curves demonstrate high threshold-independent discriminative power in distinguishing between multiple skin pathologies on the HAM10000 test set, led by the soft-voting ensemble (AUC = 0.971).

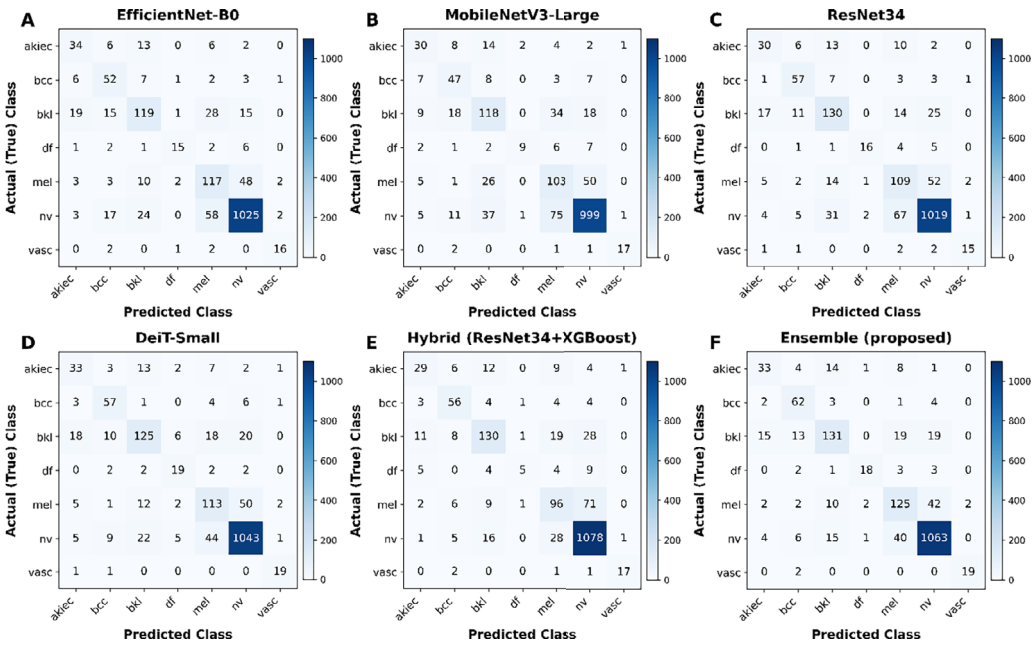

Figure 4. Detailed error analysis through confusion matrices, confirming that the soft-voting approach yields the most reliable predictions across both rare and common classes (A-F). *Abbreviations*: BKL: Benign keratosis-like lesions, NV: Benign melanocytic nevus, AKIEC: Actinic keratosis-intraepithelial carcinoma, VASC: Vascular lesions, MEL: Melanoma, BCC: Basal cell carcinoma, DF: Dermatofibroma.

class size; VASC achieved an F1-score of 0.91 despite comprising only 21 test images, whereas AKIEC (n = 61) had the lowest F1-score. The ensemble showed fewer classification errors overall than the individual architectures. In particular, MEL–NV confusion was reduced relative to ResNet34: melanoma F1 increased from 0.55 to 0.66, while the proportion of MEL cases misclassified as NV decreased from 28.1% to 22.7%. This error pattern is clinically relevant because misclassification of melanoma as a benign nevus represents a potentially consequential diagnostic error.

Cross-Validation and Statistical Comparison

Five-fold lesion-grouped cross-validation was performed for all six approaches, with the ensemble evaluated under a nested protocol as described in Methods. Mean accuracies (±SD) were: ResNet34 78.98% (±1.00), EfficientNet-B0 79.80% (±0.97), MobileNetV3-Large 78.04% (±0.84), DeiT-Small 80.21% (±2.67), the hybrid model 81.02% (±1.73), and the ensemble 82.96% (±2.07). Cross-validation accuracies were 0.2–3.1 percentage points below the corresponding test-set values across all approaches, consistent with the smaller training partitions available within each fold. The ensemble's mean cross-validated macro-F1 was 68.04% (±5.16), with the wider variability reflecting the sensitivity of macro-averaging to the small number of rare-class cases within individual folds. Fold-level results are provided in Table S3.

Differences between the ensemble and each comparator were assessed using McNemar's test, complemented by paired accuracy differences with lesion-clustered bootstrap 95% confidence intervals. The ensemble outperformed every comparator, and all confidence intervals excluded zero: MobileNetV3-Large (+7.57 pp, 95% CI 5.69–9.51; $p = 4.2\times10^{-17}$), ResNet34 (+4.43 pp, 95% CI 2.88–6.03; $p = 5.7\times10^{-9}$), EfficientNet-B0 (+4.31 pp, 95% CI 2.73–5.95; $p = 1.2\times10^{-8}$), DeiT-Small (+2.48 pp, 95% CI 1.37–3.62; $p = 3.7\times10^{-6}$), and the hybrid model (+2.36 pp, 95% CI 0.42–4.27; p = 0.006). The smallest effect was observed against the hybrid model, whose confidence interval approached zero (lower bound 0.42 pp), indicating that although the ensemble's advantage was statistically detectable, its practical magnitude over the hybrid approach was modest. Full contingency values are provided in Table S4.

Explainability and Quantitative Validation

ResNet34 was selected for detailed qualitative inspection because it achieved the highest mean Grad-CAM localization agreement with expert segmentation masks (IoU = 0.456; Table S5), rather than on the basis of classification accuracy. Grad-CAM, Grad-CAM++, and LIME visualizations were generated with respect to the model's predicted class for one representative test image from each of the seven diagnostic categories (Figure 5, panels A–G), with per-class summaries for all three convolutional architectures provided separately (Figures S2–S8). Across all classes, activation maps localized within the lesion boundaries rather than the surrounding skin, and LIME super-pixel explanations were broadly consistent with the Grad-CAM findings, identifying the lesion body as the dominant positive contributor to each prediction. Differences between Grad-CAM and Grad-CAM++ were modest for most classes, with the largest divergences observed for lesions with irregular or poorly demarcated margins.

Spatially plausible attention did not, however, guarantee correct classification. ResNet34 correctly classified five of the seven representative examples (BCC, BKL, DF, MEL, and NV) and misclassified two. The actinic keratosis example was predicted as benign keratosis (Figure 5A; Figure S2), and the vascular lesion example was predicted as actinic keratosis (Figure 5G; Figure S8). In both cases, activation remained concentrated on the lesion body despite the incorrect prediction. The actinic keratosis case is further informative: EfficientNet-B0 and MobileNetV3-Large classified it correctly while ResNet34 did not, even though all three models attended to the same central lesion region (Figure S2). These observations indicate that anatomically plausible localization is necessary but not sufficient for diagnostic accuracy, and that agreement between attention maps does not imply agreement between predictions.

Grad-CAM localization was quantitatively evaluated to move beyond qualitative visual inspection. IoU values were computed for all 1,692 test images against the HAM10000 segmentation masks, with Grad-CAM generated for the ground-truth class. Mean IoU was highest for ResNet34 (0.456 ± 0.194), followed by EfficientNet-B0 (0.427 ± 0.192) and MobileNetV3-Large (0.337 ± 0.205); per-class results are provided in Table S5. Localization agreement did not directly parallel classification performance: MobileNetV3-Large obtained both the lowest test accuracy (78.19%) and the lowest mean IoU, whereas ResNet34 ranked fifth of six in accuracy yet achieved the highest IoU among the evaluated CNNs, suggesting that classification performance and spatial localization reflect partly distinct model properties. Agreement also varied across classes; for ResNet34, mean IoU ranged from 0.486 for melanoma to 0.265 for vascular lesions. The substantial standard deviations further indicate considerable image-level variability in localization agreement. IoU-based validation was restricted to the convolutional architectures because the implemented Grad-CAM procedure was not directly applicable to DeiT-Small; this limitation is addressed in the Conclusions.

Conclusions

This study developed and quantitatively evaluated an explainable ensemble framework combining CNN- and

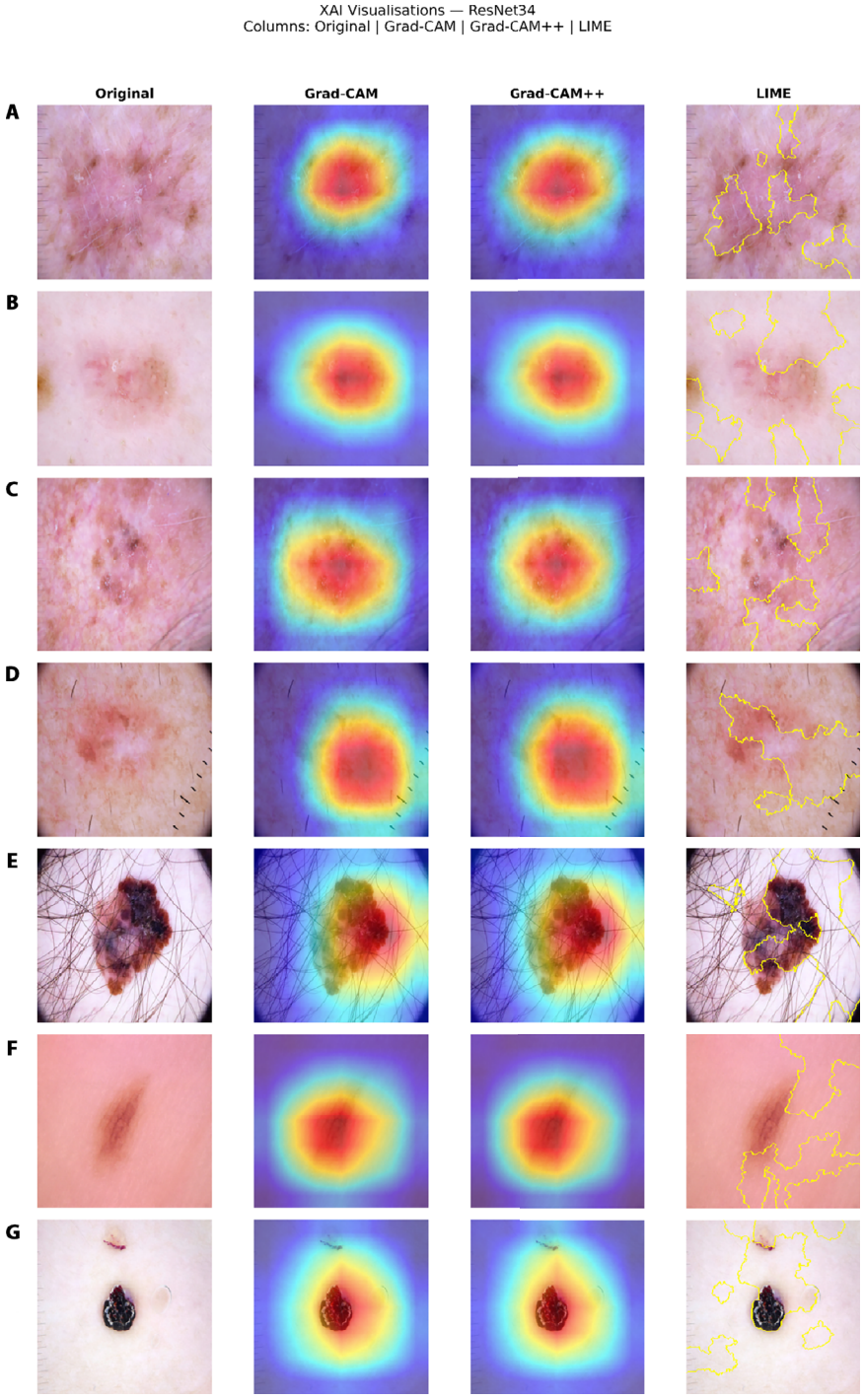

Figure 5. Grad-CAM, Grad-CAM++, and LIME visualizations generated with ResNet34 for one representative test image from each diagnostic class: (A) Actinic keratoses and intraepithelial carcinoma (AKIEC; misclassified as BKL), (B) Basal cell carcinoma (BCC), (C) Benign keratosis-like lesions (BKL), (D) Dermatofibroma (DF), (E) Melanoma (MEL), (F) Melanocytic nevi (NV), and (G) Vascular lesions (VASC; misclassified as AKIEC).

transformer-based architectures through validation-weighted soft voting for multi-class skin lesion classification on HAM10000 under a lesion-grouped partitioning protocol. The ensemble achieved 85.76% accuracy, a macro-F1 of 75.24%, and a macro-AUC of 0.971, exceeding all four individual architectures and the ResNet34+XGBoost hybrid model. Paired accuracy improvements ranged from 2.36 to 7.57 percentage points, with lesion-clustered 95% confidence intervals excluding zero in all comparisons. The ensemble achieved the highest F1-score across all seven diagnostic categories and reduced melanoma–nevus confusion relative to the individual architectures (e.g., from 28.1% to 22.7% versus ResNet34).

Direct comparison with the studies in Table 4 is complicated by differences in evaluation protocols: several do not state the partitioning unit, and at least one applies

Table 4. Comparison of the proposed model with previous HAM10000 studies, including the reported data partitioning protocol.

Study	Year	Method	XAI	Split protocol	Reported performance
Sönmez et al. [12]	2023	MobileNetV2	–	Image-level	80.79%
Mridha et al. [13]	2023	CNN	Grad-CAM, Grad-CAM++	Image-level	82.00%
Tajerian et al. [14]	2023	EfficientNet-B1	–	Image-level	84.30%
Chatterjee et al. [15]	2024	Ensemble CNN	–	Not reported	86.00%
Prasad et al. [16]	2024	EfficientNet-B3	–	Not reported	90.62%
Miętkiewicz et al. [17]	2025	DINOv2-Large (self-supervised foundation model)	Attention maps	Patient-level	Macro-F1 = 0.85
Aras & Doğan [18]	2026	MobileNet, DenseNet, Xception	Grad-CAM, LIME	Not reported	84.04%
Randhawa et al. [19]	2026	EfficientNet-B0 + MC Dropout	–	Lesion-level	Macro-F1 = 0.731
Ahmmed et al. [20]	2026	Xception + CBAM + LoRA-MedGemma	Grad-CAM++	Patient-level	85.66%
Proposed	**2026**	**EfficientNet-B0, MobileNetV3-Large, ResNet34, DeiT-Small, ResNet34+XGBoost, Ensemble**	**Grad-CAM, Grad-CAM++, LIME**	**Lesion-grouped**	**85.76% (Macro-F1 = 0.752)**

Abbreviations: XAI, explainable artificial intelligence; MC Dropout, Monte Carlo dropout. Split protocols are reported according to the information explicitly provided in the corresponding studies; "not reported" indicates that the partitioning unit was not stated. Performance values are test-set accuracy unless macro-F1 is specified. Miętkiewicz et al. additionally evaluated DermNet and ISIC Atlas; the value shown corresponds to their reported HAM10000 result.

image-level splitting after augmentation [13], potentially allowing images of the same lesion or their augmented derivatives to occur across subsets. Our findings are consistent with the range of recent lesion-aware evaluations, including 85.66% accuracy under patient-level separation [20] and a macro-F1 of 0.731 with EfficientNet-B0 [19], while a macro-F1 of 0.85 has been reported using the self-supervised DINOv2-Large foundation model [17]. The contribution of this study therefore extends beyond overall accuracy to the joint use of leakage-aware partitioning, nested cross-validation with fold-specific weight selection, class-balanced performance assessment, paired effect sizes with lesion-clustered confidence intervals, and quantitative localization analysis.

ResNet34 achieved the highest mean IoU (0.456) among the evaluated CNNs despite ranking fifth of six classifiers in accuracy, whereas MobileNetV3-Large showed both the lowest accuracy and IoU, suggesting that classification performance and spatial localization reflect partly distinct model properties. Lesion-focused activation did not necessarily correspond to correct classification: two of seven representative cases were misclassified despite activation remaining concentrated on the lesion region, and different architectures occasionally produced different predictions while attending to similar regions. Explainability outputs should therefore be interpreted as complementary to, rather than substitutes for, quantitative performance evaluation.

Class-level performance remained uneven, with AKIEC lowest (F1 = 0.564), followed by melanoma (0.656) and BKL (0.706). Although minority classes were oversampled to 6,000 samples, this did not increase the number of independent lesions, and augmentation cannot fully compensate for limited underlying diversity or inter-class similarity. Test-set support likewise did not directly determine performance: VASC achieved an F1 of 0.905 with only 21 test images, whereas AKIEC (n = 61) showed the lowest F1 and was most frequently confused with BKL.

Several limitations apply. Melanoma sensitivity was 0.676, with approximately one third of melanoma cases misclassified, indicating that sensitivity would require particular attention in future clinically oriented operating-point selection. Cross-validated macro-F1 varied more than accuracy (68.04% ± 5.16 versus 82.96% ± 2.07), and fold-specific ensemble weights varied across folds (Table S3), indicating sensitivity of the selected ensemble configuration to the training

partition. The study was evaluated on a single public dataset without external validation; no dermatologist reader study was performed, and therefore the clinical utility of the explanations was not assessed. IoU validation was also restricted to the CNN architectures because the implemented Grad-CAM procedure was not directly applicable to DeiT-Small.

Several directions follow from these findings. The most immediate is external and prospective validation, together with reader studies to determine whether the explanations are useful to clinicians beyond their spatial agreement with lesion regions. Our results also raise a further question: because classification performance and localization quality ranked the architectures differently, future work could investigate whether ensemble weighting might incorporate localization quality alongside class-balanced predictive performance. For minority classes, increasing the diversity of training examples through approaches such as generative augmentation could be explored alongside complementary strategies such as loss reweighting and class-specific threshold calibration. Extending quantitative localization analysis to transformer architectures would also allow CNN- and attention-based models to be evaluated more consistently. Finally, incorporating clinical metadata and uncertainty estimation may further improve the assessment of model reliability in real-world dermatological settings.

Authorship: BDC conceptualized and designed the study, carried out data preprocessing, model development, and evaluation, developed all software, performed the statistical analyses, and drafted the manuscript. BB critically reviewed the manuscript in detail from a dermatological perspective and made substantial revisions. EA critically reviewed the final version of the manuscript for the accuracy of clinical terminology and dermatological content. AT provided methodological guidance on the design of the model architecture and the experimental framework, supervised the study, and critically revised the manuscript for important intellectual content. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Ethics Approval and Clinical Trial Registration: Not applicable. This study used only the publicly available, fully de-identified HAM10000 dataset; no new data were collected from human participants, and therefore ethics committee approval was not required.

Consent for Publication: Not applicable.

Competing Interests: The authors declare that they have no competing interests.

Funding: No funding was received for this work.

Data Availability: The HAM10000 dataset analyzed in this study is publicly available from the Harvard Dataverse repository (https://doi.org/10.7910/DVN/DBW86T).

Use of Artificial Intelligence (AI)-Assisted Technologies: Claude Sonnet 5 (Medium) was used to assist with language polishing in the preparation of this manuscript. Following the use of this tool, the authors reviewed and edited the content thoroughly and hereby declare full responsibility for the authenticity of the contents of this publication. All scientific content, results, and conclusions remain solely the work of the authors.

Acknowledgments: None.

References

1. Dobre EG, Surcel M, Constantin C, Ilie MA, Caruntu A, Caruntu C, Neagu M. Skin cancer pathobiology at a glance: a focus on imaging techniques and their potential for improved diagnosis and surveillance in clinical cohorts. *Int J Mol Sci*. 2023;24(2):1079. DOI: 10.3390/ijms24021079
2. Debelee TG. Skin Lesion Classification and Detection Using Machine Learning Techniques: A Systematic Review. *Diagnostics*. 2023; 13(19):3147. DOI: 10.3390/diagnostics13193147
3. Arrieta AB, Rodríguez ND, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. DOI: 10.1016/j.inffus.2019.12.012
4. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalised gradient-based visual explanations for deep convolutional networks. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*; March 2018:839-847. DOI: 10.1109/WACV.2018.00097
5. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset is a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5:180161. DOI: 10.1038/sdata.2018.161
6. Akyel C, Arıcı N. Hair Removal and Lesion Segmentation with FCN8-ResNetC and Image Processing in Images of Skin Cancer. *Bilişim Teknolojileri Dergisi*. 2022;15(2):231-238. DOI: 10.17671/gazibtd.1060330
7. Ajani AA, Gambo I, Faniyi O. Artificial Intelligence-Assisted Diagnosis of Atopic Eczema in Darker Skin Types: A CNN-Based Study. *Dermatol Pract Concept*. 2026;16(1):5897. DOI: 10.5826/dpc.1601a5897
8. Sangwan H. Quantifying explainable AI methods in medical diagnosis: a study in skin cancer. *medRxiv*. Published online December 10, 2024. DOI:10.1101/2024.12.08.24318158
9. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022;12:5979. DOI: 10.1038/s41598-022-09954-8

10. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12(2):153-157. DOI: 10.1007/BF02295996
11. Fiaz M, Shoaib Khan MB, Khan AH, Bilal A, Abdullah M, Darem AA, Sarwar R. An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification. *Front Med (Lausanne)*. 2025;12:1681542. DOI: 10.3389/fmed.2025.1681542
12. Sönmez AF, Çakar S, Cerezci F, Kotan M, Delibaşoğlu İ, Çit G. Deep learning-based classification of dermoscopic images for skin lesions. SAUCIS. 2023;6(2):114-122. doi:10.35377/saucis.1314638
13. Mridha K, Uddin MM, Shin J, Khadka S, Mridha MF. An interpretable skin cancer classification using an optimised convolutional neural network for a smart healthcare system. *IEEE Access*. 2023;11:41003-41018. DOI: 10.1109/ACCESS.2023.3269694
14. Tajerian A, Kazemian M, Tajerian M, Akhavan Malayeri A. Design and validation of a new machine-learning-based diagnostic tool for the differentiation of dermatoscopic skin cancer images. PLoS One. 2023;18(4):e0284437. DOI: 10.1371/journal.pone.0284437
15. Chatterjee S, Gil JM, Byun YC. Early detection of multiclass skin lesions using a transfer learning-based IncepX-Ensemble model. *IEEE Access*. 2024;12:113677-113693. DOI: 10.1109/ACCESS.2024.3432904
16. Prasad CR, Bilveni G, Priyanka B, Susmitha C, Abhinay D, Kollem S. Skin cancer prediction using modified EfficientNet-B3 with deep transfer learning. *In: Proceedings of the 2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)*; 2024:519-523. DOI:10.1109/ICWITE59797.2024.10503193
17. Miętkiewicz Ł, Ciechanowski L, Jemielniak D. The skin game: revolutionizing standards for AI dermatology model comparison. arXiv. Preprint posted online February 4, 2025. DOI:10.48550/arXiv.2502.02500
18. Aras HH, Doğan N. A comparative analysis for skin cancer detection by using explainable deep learning. *Neural Comput Appl*. 2026;38(4). DOI: 10.1007/s00521-025-11809-y
19. Randhawa P, Suddala S, Hemal MAKP, et al. Skin lesion classification in low-resource settings using lightweight CNNs with uncertainty estimation. *BioMedInformatics*. 2026;6(4):53. DOI: 10.3390/biomedinformatics6040053
20. Ahmmed F, Antik AR, Mehnaj S, Alaly A, Mridha MF. From heatmaps to language: explainable skin cancer diagnosis using CNN–VLM integration with an expert-annotated dataset. *Artif Intell Med*. 2026;181:103499. DOI: 10.1016/j.artmed.2026.103499