5.ULUSLARARASI SELÇUK BİLİMSEL ARAŞTIRMALAR KONGRESİ 14-15 ARALIK 2024 KONYA



Editörler:

Prof. Dr. İlker Türker

Dr. Khadija Hassan Ali Al-Qusayr



KONGRE KİTABI

İKSAD Publishing House

www.isarconference.org

DEEPFAKE VIDEO DETECTION WITH RESNET50 AND ADVANCED MACHINE LEARNING ALGORITHMS

Aynur KOÇAK
Gazi University, Faculty of Technology, Department of Electrical and Electronics Engineering
ORCID ID: 0000-0001-9647-7281
Dr. Esra SÖĞÜT
Gazi University, Faculty of Technology, Department of Computer Engineering,
ORCID ID: 0000-0002-0051-2271
Prof. Dr. Mustafa ALKAN
Gazi University, Faculty of Technology, Department of Electrical and Electronics Engineering, 06560
ORCID ID: 0000-0002-9542-8039

ABSTRACT

In the digital age, the proliferation of manipulated media content threatens public trust and information integrity. The increasing resemblance between fake content and real content is becoming increasingly apparent, as the methods of producing deceptive material diversify and the emergence of increasingly sophisticated and credible products continues to grow. In this context, the detection of deepfake videos has become a critical research area for digital security and integrity. The ResNet50 model used in this study aims to capture the fine details critical for deepfake detection by extracting important features from each frame of the videos. Celeb-Df was used as the dataset in the study. In the classification phase, two powerful machine learning algorithms, Random Forest and XGBoost, were chosen. This choice provides the opportunity to compare the performance of different approaches and determine the most effective method. Another noteworthy aspect of the study is the use of the AUC metric for performance evaluation. This choice allows for a more comprehensive evaluation of the performance of the classifier at different thresholds. AUC values of 93% were obtained with Random Forest and 95% with XGBoost. The high AUC values obtained show that the proposed method is quite successful in deepfake detection. The slightly higher performance of the XGBoost algorithm highlights its superiority in modeling complex data structures.

Keywords: Deepfake video detection, ResNet50, machine learning, Random Forest, XGBoost

1. INTRODUCTION

The similarity of fake content to real content is increasing day by day, which is becoming more and more evident with the diversification of fake content production methods and the emergence of increasingly successful and credible products. In particular, the rapid advancement of deepfake technology has greatly increased the ability to realistically mimic fake media, making the need for robust and reliable detection methods even more important. With the proliferation of this

technology, research into new approaches that can effectively detect and minimize the potential risks posed by synthetic media has gained great importance (Nguyen et al. 2022). Therefore, in order to eliminate the threats that deepfake technology may pose at both individual and societal levels, it is inevitable that detection methods in this field should be continuously improved and new methods should be developed (Westerlund 2019). In line with the described challenges, this paper proposes a solution to address these problems in deepfake detection. The proposed approach aims to detect fake content with high accuracy rates despite its increasing complexity and diversity. In this paper, we present a novel approach that addresses the shortcomings of existing detection methods and adapts to the rapid advancement of deepfake technology. This approach aims to identify fake content more effectively, while at the same time contributing to a faster and more reliable minimization of the risks posed by such threats. The proposed methodology is expected to provide significant improvements over existing solutions in reducing the risks associated with synthetic media and deepfake detection.

2. RELEATED WORKS

The increase in fake content and disruptive factors have drawn attention to the methods used to detect real and fake content. Traditional methods for deepfake detection have been replaced by hybrid models or ensemble methods.

The authors propose a framework for deepfake detection based on multiple attention mechanisms. This framework identifies important local regions using attention maps and detects more complex details by enhancing texture features from superficial layers. In the final stage, low-level texture features and high-level semantic features are combined and used with the guidance of attention maps. The proposed method is applied to the Celeb-Df dataset and achieved an AUC of 67.44% (Zhao et al. 2021).

In the model proposed in (Gu et al. 2021) deepfake video detection is achieved through a temporal and spatial inconsistency learning process. This is implemented with Spatial-Temporal Inconsistency Learning (STIL) block. In this way, both temporal and spatial information are comprehensively combined and inconsistencies in the videos are captured. For this purpose, the frames in the videos are considered both horizontally and vertically. The AUC of the proposed model on the Celeb-Df dataset is obtained as 75%.

In (Zhang et al. 2021), the authors proposed the Temporal Dropout 3-Dimensional Convolutional Neural Network (TD-3DCNN) method. In the proposed method, fake video detection is performed by utilizing clues arising from the inconsistency between frames in videos. It extracts features by feeding fixed-length frames into 3D Convolutional Neural Network and uses these features for classification. The model proposed by the authors was tested on the Celeb-Df dataset and achieved

an AUC of 88%. Face morphing techniques manipulate the facial region to make it congruent with its context, while leaving the surrounding context intact. This results in mismatches between the face and the context, and these mismatches provide clues for the detection of manipulation.

In (Nirkin et al. 2021), a detection model is developed by addressing these inconsistencies. In the proposed method, a face recognition network is used to analyze the facial region and a context recognition network to evaluate facial contexts. Manipulations are detected with the signals provided by these two networks. The AUC value achieved by the proposed method is 66%.

Another work (Gu et al. 2022) that studies temporal inconsistencies proposes the Hierarchical Contrastive Inconsistency Learning (HCIL) framework. Unlike the others, this work introduces a two-stage contrast learning method to detect temporal inconsistencies in a more detailed and comprehensive way at both local and global levels. The HCIL model includes a region-adaptive module for more efficient detection of inconsistencies between video segments and a fusion module for information fusion between segments. For local contrast learning, a region-aware discrepancy module is designed that generates dynamic video segment representations and an innovative weighted Noise Contrast Estimation (NCE) loss. Global contrast learning is performed using a fusion module to combine the information between segments. This method achieved an AUC of 79% on the Celeb-Df dataset.

The authors introduced a new deepfake forensic detection method called “forensic symmetry” (Li et al. 2023). This approach aims to determine whether two symmetric parts of the face have the same or different natural features. The main goal of the work is to detect deepfake manipulations by examining the consistency of natural features in symmetric regions of the face. The method provides a more sensitive and specific detection method by focusing on specific regions of the face. This proposed approach achieved an AUC of 61.97% on the Celeb-Df dataset.

The SIGMA-DF method developed in this paper (Han et al. 2023) presents an approach to address cross-domain issues in deepfake detection. Instead of directly extracting forgery traces, SIGMA-DF uses a meta-learning strategy to simulate transitions between different domains during training. This allows the model to better generalize information from different domains. Ensemble meta-learning and the use of a new weighted single-side loss function help the model to discriminate between real and fake faces. Testing the proposed method on the Celeb-Df dataset by the authors, the AUC value obtained is 91.3%.

The study (Ye et al. 2024) also proposes a solution to the difficulties in the generalization capability of detection models through semantic decomposition. The proposed method emphasizes the separation of common fraud semantics and the use of these semantics to improve the overall detection capabilities of the models. An adaptive high-pass module and a two-stage training

strategy are developed and their performance is demonstrated. The AUC value achieved by the method on the Celeb-Df dataset is 76.94%.

The method proposed in (Lin et al. 2024) addresses the issue of fairness preservation in deepfake detection. The paper presents a method for deepfake detection that improves the overall validity while preserving fairness across different demographic groups. The proposed approach provides a fair learning process by using decomposition learning to extract demographic and domain-agnostic forgery features and combining these features. Furthermore, a loss surface smoothing strategy is added, making it easier to optimize these modules. The AUC achieved by the proposed method is 78.55%.

3. PROPOSED MODEL

The stages and general information of the model proposed in the study were presented in this section. The flow diagram of the proposed model was shown in Figure 1. Accordingly, the determined stages were carried out in order.

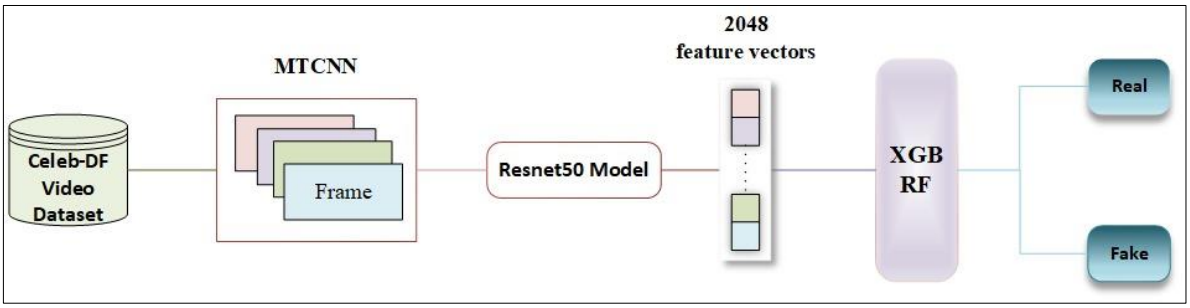


Figure 1. Flow chart of the proposed model

3.1. Dataset

Celeb-Df is a comprehensive dataset of fake videos created using deep learning-based face manipulation techniques. The dataset presents realistic videos that are widely used in studies on deepfake detection, increasing the challenges of the field (Li et al. 2020). There are 590 real and 5639 fake videos in the dataset. The real videos consist of videos of 59 celebrities posted on YouTube. The fake videos are characterized by natural facial expressions, lip synchronization and other facial details presented with a high degree of believability. The fake videos were created using FaceSwap (FaceSwap 2024).

3.2. Frame extraction

Multi-task Cascaded Convolutional Network (MTCNN), is mission critical in the field of face detection, computer vision and is used as a core component in many applications (Zhang et al. 2016). In this study, MTCNN uses a three-stage network structure for face detection and landmark points detection. MTCNN has a three-stage network structure that simultaneously performs the

tasks of face detection and landmark detection: P-Net, R-Net and O-Net. These networks perform face detection on the image.

MTCNN offers high accuracy and efficiency in face detection and landmark detection tasks. The progressive structure enables the detection of faces of different sizes and the accurate identification of critical points on the face. The multi-tasking nature of MTCNN makes it an ideal solution for a variety of applications.

3.3. Feature extraction

Deep learning models have made great advances in classification and image recognition tasks. However, as the network depth increases, problems such as vanishing gradient arise during training. ResNet50 uses the concept of residual learning to overcome the main challenges in training deep neural networks (He et al. 2016). This model effectively solves the problem of gradient fading while improving the performance of deep networks.

The basic building blocks of ResNet50 are layers called residual blocks. Each residual block contains a set of convolutional layers, and the output of these layers is added to the input. These direct connections enable the network to be trained more extensively and enhance the model's overall effectiveness. In contrast to traditional deep networks, residual blocks provide better propagation of gradients and thus prevent performance degradation during training.

3.4. Classification

The use of machine learning methods in classification problems offers the ability to perform in-depth analysis on large and complex datasets. These methods have the capacity to automatically learn complex patterns and relationships in datasets that are difficult to perceive by humans. In particular, these algorithms can classify structured and unstructured data types with high accuracy and offer broad generalizability. Furthermore, with the ability to continuously retrain themselves with new data, machine learning models optimize their performance over time, which improves classification accuracy.

XGBoost is a robust machine learning algorithm initially introduced by Chen Tianqi and Carlos in 2011. Since its inception, it has seen ongoing enhancements and refinements from numerous researchers (Li et al. 2019). The model is known for its high accuracy and efficiency, especially on different data types and complex data structures (Maulana et al. 2023). XGBoost builds decision trees sequentially, and each new tree iteratively improves prediction performance by correcting the errors of previous models. This iterative process allows XGBoost to make more accurate predictions over time, so the algorithm achieves high performance on a wide range of machine learning tasks (Noviandy et al. 2023).

The goal of the XGBoost algorithm is to optimize an objective function that minimizes the total loss and model complexity. The objective function consists of two main components, the loss function ($\mathcal{L}$) and the regularization term (Ω). During the t th iteration, the objective function is defined as:

$$\mathcal{L}^{(t)} \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (1)$$

In this formula, y_i, represents the true value of the $i-th$ sample. $\hat{y}_i^{(t-1)}$ is the model's prediction for the $i-th$ sample before the $t-th$ iteration. $f_t(x_i)$, is the new decision tree added at the $t-th$ iteration. l is the loss function, which is the error between the observed value y_i and the predicted value $\hat{y}_i^{(t-1)} + f_t(x_i)$. $\Omega(f_t)$ is the regularization term that controls the complexity of the model at the $t-th$ iteration.

The objective function increases the accuracy of the model while at the same time preventing overfitting.

$$\Omega(\mathrm{f}) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \quad (2)$$

In the given formula, T, denotes the number of leaves of the decision tree. γ, is the penalty coefficient controlling the model complexity. λ is a regularization parameter based on the L_2 norm. $\|w\|^2$ is the L_2 norm of the vector containing the scores of the leaves.

Random Forest (RF) is a machine learning algorithm that leverages an ensemble of decision trees and is commonly applied to classification and regression tasks (Zhang et al. 2012). The fundamental approach of this algorithm involves training multiple decision trees independently, with the final prediction being derived from aggregating the outputs of these trees. The dataset D, contains a total of η observations and m features. In the first step, k features are randomly selected from the dataset ($k \leq m$). Each decision tree is trained with a sampled subset and k features. This ensures that each tree learns on a different subset and feature combination, thus providing model diversity. In classification problems, each decision tree produces predictions and the final decision is made according to the majority of these predictions (Ogutu et al. 2011).

4. EXPERIMENTS

4.1. Experimental Settings

The Celeb-Df dataset has been widely used as a benchmark for the detection and classification of realistic and high-quality fake video content in studies on deepfake video detection. Moreover, the wide diversity and high level of realism contained in the dataset encourages algorithms to adapt to different scenarios, not only to a specific type of fake video. Therefore, Celeb-Df is an ideal source for measuring the effectiveness of algorithms used in deepfake detection and laying the

groundwork for the development of new methods. In order to balance the number of real and fake videos, we used all 590 real videos and about one-fifth of the fake videos in the dataset.

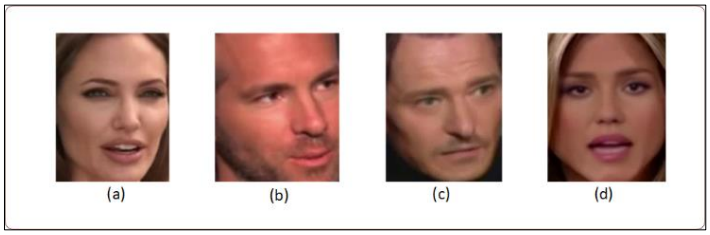


Figure 2. Real and fake frames, (a) and (b) real images, (c) and (d) fake images

The MTCNN method was used for face detection and frame extraction from the videos in the Celeb-Df dataset. For each video, 4 to 5 frames were selected at certain intervals and the faces in these frames were successfully detected and processed. This process preserved the representativeness of the videos, transformed the dataset into an analyzable form, and created a dataset that can be used in the training and testing processes of the model for deepfake detection. Figure 2 provides examples of real and fake images in the dataset.

The total number of images obtained from real and fake videos is 7,724 (Table 1). The images are divided into a 70%-30% training and test dataset. This approach aims to increase the efficiency of the analysis process while maintaining the diversity of the dataset.

Table 1. Number of frames

	Train	**Test**	**Total Frame**
Real Frame	1,856	848	2,704
Fake Frame	3,550	1,470	5,020
Total Frame	5,406	2,318	7,724

4.2. Pre-processing

In the preprocessing step, image processing operations were applied on the frames extracted from the videos. First, each frame was read using the OpenCV library and then resized to the target size of 299x299 pixels. The images were converted from BGR color space to RGB color space and normalized according to the requirements of the model. The normalization process involved rescaling each pixel value from 0 to 1 and then standardizing it using the mean and standard deviation values (μ=[0.485,0.456,0.406], σ=[0.229,0.224,0.225]). These steps were performed in order to obtain more consistent and generalizable results during the training of the model.

4.3. ResNet50

After preprocessing, the ResNet50 model was employed to extract deep features from the images. ResNet50 is a convolutional neural network based on deep learning that efficiently captures high-

level features from images due to its layered architecture. In this study, a pre-trained ResNet50 model was utilized by removing its final layer, allowing feature maps to be extracted from each frame. These feature maps are then used as input during the classification phase to enhance the model's performance in detecting deepfakes.

4.4. Machine learning

After feature extraction, classification is performed for deepfake detection. For this purpose, RF and XGBoost algorithms were used. RF is an ensemble method that works as an ensemble of multiple decision trees and determines the final classification by majority voting or averaging the independent predictions of each tree. XGBoost is a powerful ensemble learning algorithm based on gradient boosting and is known for its high accuracy rates. These two algorithms are used to efficiently classify the extracted features to optimize the model's performance in deepfake detection.

Area Under the Curve (AUC) represents the area under the ROC curve and summarizes the performance of the model at different classification thresholds. This objectively reflects how successfully the model distinguishes between both positive and negative classes, especially in datasets where class imbalance is present. Furthermore, because the AUC metric can evaluate independently of a given decision threshold, it provides a reliable way to measure the overall performance of the model, even when different thresholds are used in different applications. For these reasons, AUC is an ideal metric to highlight model performance in deepfake detection in a comprehensive and understandable way. Figure 3 shows the AUC values of the two algorithms used for classification.

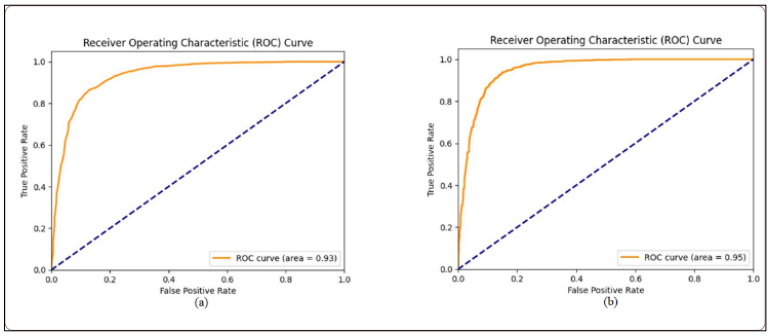


Figure 3. AUC values of the proposed models, (a) RF algorithm, (b) XGBoost algorithm

Table 2 shows the AUC values of the studies that detect deepfake videos using Celeb-Df dataset and the proposed method.

Table 2. AUC values of the studies

References	AUC
(Zhao et al. 2021)	%67.44
(Gu et al. 2021)	%75.00
(Zhang et al. 2021)	%88.00
(Nirkin et al. 2021)	%66.00
(Gu et al. 2022)	%79.00
(Li et al. 2023)	%61.97
(Han et al. 2023)	%91.30
(Ye et al. 2024)	%76.94
(Lin et al. 2024)	%78.55
Our Study	**%95**

5. Conclusions

This study aims to develop an effective method for the detection of deepfake videos, which is an increasing threat. The rapid development and proliferation of deepfake technology makes it difficult to distinguish between real and fake content, which can lead to serious problems such as information manipulation and misleading propaganda. Therefore, the development of reliable and high-performance deepfake detection methods is vital for information security in the digital age. In the first step, key frames were extracted from the analysed videos. This is an important pre-processing step to analyse the static visual features of the video content. ResNet50 model is used to extract feature vectors from the extracted frames. RF and XGBoost are used for classification to compare the performance of both traditional and modern machine learning approaches. Since it is a reliable metric that measures the discrimination power of the model in classification problems, performance evaluation was performed with the AUC metric. The 93% with the RF model and 95% with the XGBoost method show that the developed method is very successful in detecting deepfake videos. The slightly higher performance of XGBoost emphasizes the superiority of this algorithm in modelling complex data structures.

REFERENCES

Nguyen TT, Nguyen QVH, Nguyen DT, Nguyen DT, Huynh-The T, Nahavandi S, et al. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, vol. 223, p. 103525.

Westerlund M. (2019). The emergence of deepfake technology: A review. Technology innovation management review, vol. 9.

ZhaoH, Zhou W, Chen D, Wei T, Zhang W, and Yu N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2185-2194.

Gu Z, Chen Y, Yao T, Ding S, Li J, Huang F, et al. (2021). Spatiotemporal inconsistency learning for deepfake video detection. In Proceedings of the 29th ACM international conference on multimedia, pp. 3473-3481.

Zhang D, Li C, Lin F, Zeng D, and Ge S. (2021). Detecting Deepfake Videos with Temporal Dropout 3DCNN. In IJCAI, pp. 1288-1294.

Nirkin Y, Wolf L, Keller Y, and Hassner T. (2021). Deepfake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, pp. 6111-6121.

Gu Z, Yao T, Chen Y, Ding S, and Ma L. (2022). Hierarchical contrastive inconsistency learning for deepfake video detection. In European Conference on Computer Vision, pp. 596-613.

Li G, Zhao X, and Cao Y. (2023). Forensic symmetry for deepfakes. IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1095-1110.

Han B, Li J, Ren W, Luo M, Liu J, and Cao X. (2023). SIGMA-DF: Single-Side Guided Meta-Learning for Deepfake Detection. In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, pp. 153-161.

Ye W, He X, and Ding F. (2024). Decoupling Forgery Semantics for Generalizable Deepfake Detection. ArXiv preprint arXiv:2406.09739.

Lin L, He X, Ju Y, Wang X, Ding F, and Hu S. (2024). Preserving fairness generalization in deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16815-16825.

Li Y, Yang X, Sun P, Qi H, and Lyu S. (2020). Celeb-Df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3207-3216.

(12.06.2024). FaceSwap. Available: https://www. faceswap.dev/

Zhang K, Zhang Z, Li Z, and Qiao Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, vol. 23, pp. 1499-1503.

He K, Zhang X, Ren S, and Sun J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

Li W, Yin Y, Quan X, and Zhang H. (2019). Gene expression value prediction based on XGBoost algorithm. Frontiers in genetics, vol. 10, p. 1077.

Maulana A, Faisal FR, Noviandy TR, Rizkia T, Idroes GM, Tallei TE, et al. (2023). Machine learning approach for diabetes detection using fine-tuned XGBoost algorithm. Infolitika Journal of Data Science, vol. 1, pp. 1-7.

Noviandy TR, Idroes GM, Maulana A, Hardi I, Ringga ES, and Idroes R. (2023). Credit card fraud detection for contemporary financial management using xgboost-driven machine learning and data augmentation techniques. Indatu Journal of Management and Accounting, vol. 1, pp. 29-35.

Zhang C, and Ma Y. (2012). Ensemble machine learning, vol. 144: Springer.

Ogutu JO, Piepho HP, and Schulz-Streeck T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. In BMC proceedings, vol. 5, pp. 1-5.