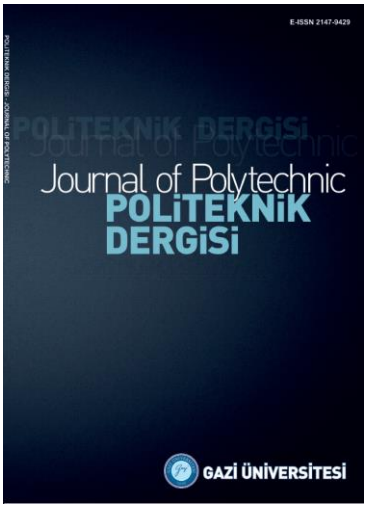

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Deepfake video detection using convolutional neural network based hybrid approach

Evrişimsel sinir ağı tabanlı hibrit yaklaşım kullanılarak deepfake video algılama

***Yazar(lar) (Author(s))**: Aynur KOÇAK[1], Mustafa ALKAN[2], Süleyman Muhammed ARIKAN[3]*

ORCID[1]: 0000-0001-9647-7281

ORCID[2]: 0000-0002-9542-8039

ORCID[3]: 0000-0003-1526-2970

To cite to this article: Koçak A., Alkan M. And Arıkan M. S., "Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Koçak A., Alkan M. And Arıkan S. M., "Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1523983

Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach

Highlights

- *A hybrid model is presented with deep learning architectures and mechanical engineering in deepfake video detection.*
- *The study used two different feature extraction methods and a total of eight hybrid models were proposed with four machine developments.*
- *The developments used the dataset frequently used in the literature and high accuracy and area under the curve (AUC) values were observed compared to other processes.*

Graphical Abstract

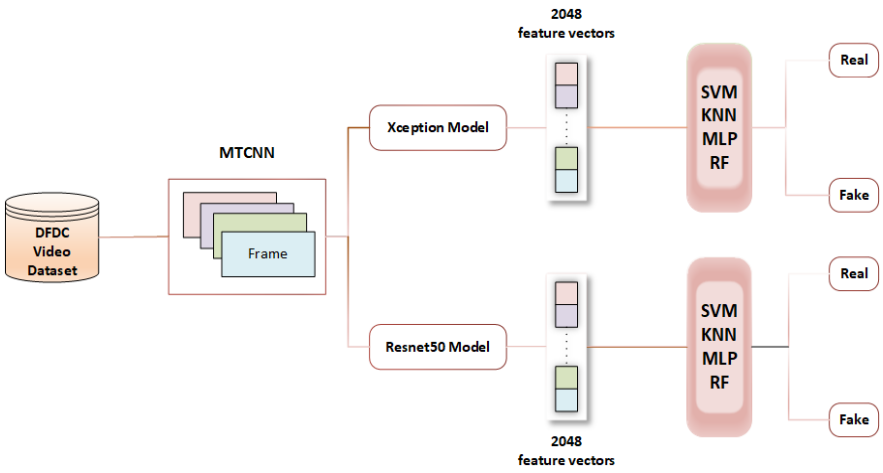

Figure. *Deepfake video detection using hybrid model*

Aim

In this study, deepfake video detection was aimed with deep learning and machine learning.

Design & Methodology

Frames were extracted from the video dataset and a feature file was created from the obtained images. Xception and ResNet50 models were used for feature extraction. The obtained feature vectors were subjected to machine learning algorithms to detect real-fake.

Originality

The performance metrics used to prove the accuracy of the study are given in detail and supported by visuals. In addition, the performance metric values obtained have managed to exceed the results in the literature.

Findings

The accuracy (ACC) and AUC values achieved after classification achieved high success in deepfake video detection.

Conclusion

As a result; After feature extraction with Xception and ResNet50, deepfake video detection was successfully completed with the classification methods applied. Recently, a study has been presented to detect deepfake technology used for malicious purposes throughout the country.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Deepfake Video Detection Using Convolutional Neural Network Based Hybrid Approach

Araştırma Makalesi / Research Article

Aynur KOÇAK[1*], Mustafa ALKAN[1], Muhammed Süleyman ARIKAN[2]

[1]Faculty of technology, Department of Electrical and Electronics Engineering, Gazi University, Turkey
[2] Institute of Informatics, Computer Forensics, Gazi University, Turkey

(Geliş/Received : 29.07.2024 ; Kabul/Accepted : 03.09.2024 ; Erken Görünüm/Early View : 22.11.2024)

ABSTRACT

Given the rapid advancement of deepfake technology, which allows for the creation of highly realistic fake content, there is a pressing need for an efficient solution to address the security risks associated with this technology. Deepfake videos are widely recognized for their significant implications, including the potential for identity theft, the dissemination of false information, and the endangerment of national security. Therefore, it is crucial to develop and enhance the reliability of deepfake detection algorithms. In this study, feature extraction techniques were performed to utilize deep learning algorithms such as Xception and ResNet50 to detect deepfakes in a video dataset using the DFDC dataset. Additionally, a total of eight hybrid models were developed using various classification algorithms such as SVM, KNN, MLP, and RF. The ResNet50 and RF hybrid models achieved the highest accuracy rate of 98%, with an AUC value of 99.65%. This study presents a machine learning method that has been developed to address different technical challenges in the field of deepfake detection and effectively identify deepfakes. The proposed method has demonstrated successful performance compared to state-of-the-art models, proving its effectiveness in accurately detecting fake content within videos.

Keywords: Deepfake video detection, deep learning, machine learning, Xception, ResNet50.

Evrişimsel Sinir Ağı Tabanlı Hibrit Yaklaşım Kullanılarak Deepfake Video Algılama

ÖZ

Son derece gerçekçi sahte içeriklerin oluşturulmasına olanak tanıyan deepfake teknolojisinin hızla ilerlemesi göz önüne alındığında, bu teknolojiyle ilişkili güvenlik risklerini ele almak için etkili bir çözüme acil ihtiyaç duyulmaktadır. Deepfake videoları, kimlik hırsızlığı potansiyeli, yanlış bilginin yayılması ve ulusal güvenliğin tehlikeye atılması gibi önemli etkileri nedeniyle yaygın olarak bilinmektedir. Bu nedenle, deepfake tespit algoritmalarının geliştirilmesi ve güvenilirliğinin artırılması hayati önem taşımaktadır. Bu çalışmada, DFDC veri setini kullanarak bir video veri setindeki deepfake'leri tespit etmek için Xception ve ResNet50 gibi derin öğrenme algoritmalarını kullanmak üzere özellik çıkarma teknikleri gerçekleştirildi. Ek olarak, SVM, KNN, MLP ve RF gibi çeşitli sınıflandırma algoritmaları kullanılarak toplam sekiz hibrit model geliştirildi. ResNet50 ve RF hibrit modelleri, %99,65'lik bir AUC değeriyle %98'lik en yüksek doğruluk oranına ulaştı. Bu çalışma, deepfake tespiti alanındaki farklı teknik zorlukları ele almak ve deepfake'leri etkili bir şekilde tespit etmek için geliştirilen bir makine öğrenimi yöntemini sunmaktadır. Önerilen yöntem, videolardaki sahte içeriği doğru bir şekilde tespit etmede etkinliğini kanıtlayarak, mevcut modellerle karşılaştırıldığında başarılı bir performans göstermiştir.

Anahtar Kelimeler: Deepfake video tespiti, derin öğrenme, makine öğrenmesi, Xception, ResNet50.

1. INTRODUCTION

The issue of manipulated videos featuring altered faces has acquired widespread attention, particularly in the last years [1, 2]. Although deepfake technology has potential positive applications in fields like filmmaking and virtual reality, it is predominantly utilized for malicious purposes [3]. The first instance of deepfake content was a collection of pornographic videos depicting celebrities, created by a Reddit user named “deepfakes” in 2017. This indicates that the malicious use of deepfake technology was inevitable since its creation. Subsequently, various applications such as FaceApp [4], FaceSwap[5], and other tools based on deepfake technology emerged continuously. These tools allow users to modify their facial appearance, hairstyle, gender, age and other personal attributes.

The deepfake algorithm, utilizing either Auto Encoder (AE) or Generative Adversarial Network (GAN), has the capability to replace faces in target videos with faces extracted from source videos. The DeepFake AutoEncoder (DFAE) [6] is fundamentally a synthetic data generation model that employs the autoencoder method. The acronym DFAE is derived from its purpose of generating deepfake data. The method relies on the operational principles of encoder and decoder structures. It aims to generate an output image that closely resembles the input image in terms of its features. Contrarily, the GAN model utilizes two neural networks to generate

*Sorumlu Yazar (Corresponding Author)
e-posta : aynurkocak@gazi.edu.tr

counterfeit videos: (i) a generative network and (ii) a discriminative network. The generative network produces fake images from random input data. Subsequently, the discriminative network attempts to evaluate the authenticity of these generated images. As the generative network persistently strives to create increasingly realistic images in order to deceive the discriminative network, the discriminative network enhances its ability to detect fake images. This competitive process ultimately leads to the creation of highly realistic fake images. The combination of these two networks is called a Generative Adversarial Network (GAN), as proposed by Ian Goodfellow [7].

The videos generated by these applications are progressively being utilized not just to violate personal privacy, but also to interfere with political campaigns and public opinion. Identifying deepfake content is of great importance for everyone on a global scale. The increasing interest in this technology, there is a rising amount of research being carried out in this field [8].

In this study, a model was developed using Deep Learning (DL) architectures and Machine Learning (ML) to detect fake videos generated through facial manipulation. During the development phase, it is crucial to precisely define the problem, the desired output, the data type and size, along with the number of features in the data. Within this framework, a model was designed considering the size of the training data, the accuracy of the outputs and their interpretability.

2. RELEATED STUDIES

The detection of deepfake videos has rapidly emerged as a prominent topic due to the progress of Artificial Intelligence (AI) and DL techniques. These methods and technologies have facilitated the creation of realistic fake videos and their use for various purposes. Consequently, the detection and prevention of deepfake videos have also become a research subject encompassing various methodologies and techniques.

The integration of DL and AI techniques with traditional video analysis methods provides innovative and impressive solutions for detecting realistic fake videos. These methods primarily analyze the video content by examining the individual's facial features, gestures, facial expressions, or vocal intonations. Though this analysis, they can identify the differences between real and fake videos.

Chang et al. [9] utilized Python OpenCV library to extract frames in their 2020 study. For feature extraction, they based their work on the VGG [10] Network. An SRM filter layer and an image augmentation layer were incorporated before the VGG16 network. With these layers, the NA-VGG network was designed. The proposed models was evaluated on the Celeb-DF [11] dataset and obtained an Area Under the Curve (AUC) of 85.7%.

In another study conducted in the same year [12], the Multi-Task Cascaded CNN (MTCNN) [13] face detector method was utilized for image extraction from frames. The XceptionNet [14] method, pre-trained with ImageNet [15], was employed for feature extraction. Subsequently, a 2D and 3D Convolutional Neural Network (CNN) detection model was proposed to differentiate between authentic and fake videos. The 2DCNN method was utilized for image detection, whereas the 3DCNN method was employed for video detection. The researchers then applied the designed methods to their own created WildDeepFake [12] dataset, as well as the DF-TIMIT [16] and DFD [17] datasets.

In a study using the Convolutional Vision Transformer (CViT) for fake video detection, the authors combined the Vision Transformer (ViT) with the CNN model to create the CViT architecture [18]. The study comprises two components: CNN for feature extraction and Vit for categorizing these features using an attention mechanism. The Softmax function is utilized on the last layer of the CViT model in order to carry out classification. To validate the accuracy of the test result, the loss function was calculated. DFDC [6] data set was used to test the proposed model and a success rate of 91.5% was achieved. In addition to the accuracy value, a loss value of 0.32 and an AUC value of 0.91 were achieved.

In [19], the detection of fake videos was accomplished using an unsupervised comparative learning method. In this study, Xception method was employed as the frame backbone for feature engineering. The SVM algorithm utilized in the classification stage. FF++ [20], Celeb-DF and UADFV [21] datasets were used for model evaluation.

In another study, Chen et al. [22] utilized the Celeb-DF, DFDC, and FaceForensics-1.0 datasets for fake video detection. The Xception architecture was employed to extract features from both the real and the fake videos in the datasets, with certain parameters in the layers being altered. This study differs from related studies in the field in that it includes a comparative analysis. This algorithm checks whether the selected image and the reference image are the same, and if they are not, a fake of the input image is generated. The methods used for fake content generation include DeepFakes (DF), Face2Face, FaceSwap, and NeuralTextures. Following these procedures, the two images are inputted into a discriminator, which carries out the classification of the images.

In the study conducted in [23], the proposed model was applied to both their own SR-DF dataset and the Celeb-DF dataset. The proposed model utilized the Multi-Scale Transformer technique to extract the Red-Green-Blue (RGB) characteristics of the image. In addition, the Frequency Filter method was employed to extract the frequency information from the image. The extracted values were fed into the ForgeryNet network for

detection. The AUC value was obtained as 86.7 for the SR-DF dataset and 95.5 for the Celeb-DF dataset.

In the study given in [24], the Python cv2 library was used for frame extraction purposes. The ResNet and LSTM architectures were employed for feature extraction from the obtained images. The ResNet architecture was utilized for training the model to extract features, while the LSTM structure was employed for performing the classification. The method suggested by the authors was evaluated with the Celeb-DF dataset. As a result of the testing processes, a 91% accuracy rate was achieved.

In the study utilizing four different datasets for testing purposes [25], source camera noise features were analyzed across all four datasets. The utilized datasets include FF++, Celeb-DF, DFD, and DeeperForensics-1.0 [26]. The InceptionV3 architecture was employed to extract noise features. This architecture, developed by Google researchers, is used for classification and image recognition and includes numerous CNN layers. Following the capture of features with this architecture, the Siamese network is employed for classification. The proposed model was not suitable for testing with Celeb-DF dataset. However, the separation model achieved AUC values of 99.9%, 96.08% and 89.2% on the FF++, DFD and DFDC datasets, respectively.

In the study conducted in [27], the ViT and EfficientNet-B7 models were combined. Additionally, Data-Efficient Image Transformers (DeiT) were implemented for the purpose of deepfake detection. The combination of the proposed methods was tested on the Celeb-DF and DFDC datasets. As a result of the testing phase, The AUC value obtained after the testing phase for the DFDC and Celeb-DF datasets were, 0.978 and 0.993, respectively.

Yang et al. [28], utilized Spatio Temporal Attention (STA) to extract features from images taken from videos. In addition, they employed Masked Relation Learning (MRL) for the purpose of feature learning. To reveal and detect irregularities in the video, a Temporal Convolution Network (TGCN) was utilized. The proposed method was tested with FF++, Celeb-DF and DFDC datasets. It achieved accuracy rates of 98.27%, 99.96% and 99.11%, respectively.

The study conducted in [29] utilized Graph Neural Network (GNN) to transform the nodes and edges of an image into a graph, employing visual-to-visual placements. Each face frame is divided into parts using patches. In the following step, each part's graph neural network is obtained to be in the classification layer. The classification layer comprises Conv2d, Batch Normalization, PReLU, and Dropout layers. The efficiency of the proposed method was evaluated by conducting tests on various datasets including FF++, Celeb-DF, DFDC, World Leaders Dataset, and Cross Dataset.

The study presented in [30] focuses on detecting fake videos using unsupervised learning. The approach uniquely leverages fluctuations in the image. Photo-Response Non-Uniformity (PRNU) is described as the blemish caused by fluctuations from the camera or sensors. The process of extracting PRNU involves calculating noise residues and recording the corresponding PRNI values. The datasets utilized for detecting fake videos in this research are FF++, DFDC, and DFD.

In the study by Mitra et al. [31], the frame extraction method known as key frame extraction was utilized. In the next step, dlib's 68 landmark method was employed for face detection. Three methods, namely, Xception, ResNet50, and InceptionV3, were employed for feature extraction in the context of fake video detection. These methods were applied to the FF++ and DFDC datasets. Upon comparing the results, it was observed that the Xception architecture provided more successful results.

In the study conducted in 2023 [32], I-frames were extracted from videos. The researchers preferred processing on I-frames due to their higher retention of color information. The MTCNN method was employed to extract facial regions from the frames. The Xception architecture was utilized for learning features from the frames. For the feature selection step, authors proposed the Hybrid Feature Selection (HFS), which utilized the Grey Wolf Optimizer (GWO) [33] and, Vortex Search (VS) [34]. Both algorithms are classified as metaheuristic algorithms. Following the feature selection phase, the MLP method was employed for classification. The datasets used in this approach include FF++, Celeb-DF, and DFDC. The accuracy values achieved by the proposed method in the data sets are 98.00%, 97.3% and 75.34%, respectively.

In the study presented in [35], 32 frames were extracted from each video, resulting in a total of 3.8 million frames. These frames were then used for fake video detection. The MTCNN method was used for face extraction from the frames. Authors utilized EfficientNet for feature extraction from the faces and model training. During the training phase, loss function and optimization function were employed. The DFDC dataset was used for training and classification purposes, utilizing a CNN model for the classification. Authors stated that the proposed model achieved an AUC value of 92, and the minimum loss was calculated as 0.40.

Data sets frequently used in literature and academic studies related to this field, and detailed information about the contents of these data sets are given in Table 1.

Table 1. Number of videos in deepfake datasets

Dataset	Fake Video	Real Video	Total Video
DFDC [6]	104, 500	23,654	128,154
FF++ [20]	4,000	1,000	5,000
Celeb-DF [11]	5639	590	6,229
DeeperForensics-1.0 [26]	10,000	50,000	60,000

3. PROPOSED METHODOLOGY

The exponential growth of digital media in recent years has resulted in a surge in fake video content, namely deepfake videos [36]. Deepfake videos are synthetic videos generated via the use of AI and ML techniques, making them challenging to distinguish from genuine people. In this study, the DFDC dataset was chosen to be utilized for detecting deepfake videos. The MTCNN model was utilized to extract frames from the videos. The Xception and ResNet50 models were preferred for feature extraction, while various ML models such as Random Forest (RF), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP) and K-Nearest Neighbors (KNN) were employed to classify real and fake videos. This study provides a substantial contribution to the methods used in combating deepfake videos, hence helping to prevent the spread of fake content.

3.1. Dataset

This study utilizes the DFDC dataset, which provides a diverse collection of videos with various facial features, poses, backgrounds, and lighting conditions. The large quantity of videos in the dataset has a crucial role in model training and testing. The DFDC dataset contains fake videos created using different generation methods [4, 5, 37-39], which enhances the detection models' ability to identify different types of fake videos. The dataset is widely utilized and endorsed by a large research community. The dataset has a total size of 471.84 gigabytes, organized into fifty folders. Each folder has an approximate size of 10 gigabytes, and the videos they contain are 10 seconds long.

3.2. Frame Extraction

One of the crucial steps in deepfake video detection is accurately identifying the faces in the videos and extracting the frames of these faces. In our study, MTCNN is utilized to extract faces from both real and fake videos [13]. This method is specifically designed to detect faces and identify specific points on the face in a swift and accurate manner. The algorithm consists of three components, namely, the proposal network, the refine network, and the output network.

1. Proposal Network: This stage detects potential facial regions in the image. It marks possible face regions in low-resolution images and forwards them to the refine network for in-depth analysis.
2. Refine Network: This network analyzes the face regions identified by the proposal network more effectiveness and precision. It eliminates non-face regions, thereby increasing the accuracy of the face regions.
3. Output Network: This final component identifies the specific facial features within the detected area of the face. This network precisely determines the regions such as the eyes, nose, and mouth, ensuring sharper positioning of the face bounding boxes.

Accurately extracted face frames allow DL algorithms such as Xception and ResNet50, which are used in the subsequent stages, to operate more efficiently and accurately. This, in turn, makes the detection of deepfake videos using MTCNN method more reliable.

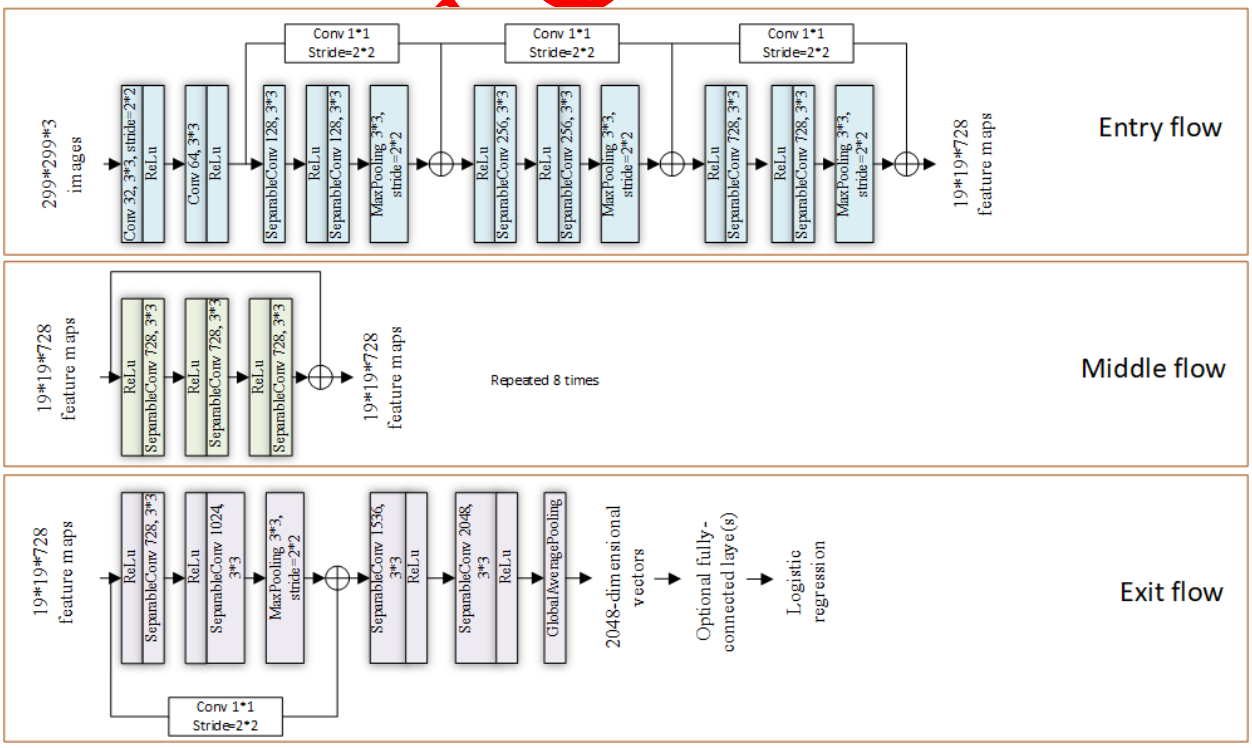

Figure 1. Xception architecture

3.3. Base Models

In the literature, frequently used algorithms such as Xception and ResNet50 have been combined with various ML algorithms to produce different models.

Xception [14] is a DL architecture developed by François Chollet in 2017, based on the CNN algorithm. Xception is particularly employed for the task of image classification. The architecture consists of multiple interconnected convolution layers, fully connected layers, pooling layers and activation functions of various types and values. The objective is to convert an input image with dimensions 229x229x3 into a feature vector with 2048 dimensions. The primary purpose of the input, middle and output flow is to generate the most appropriate vector by utilizing the most effective features of the input image. In the last layer of the architecture, the Logistic Regression algorithm is utilized for image classification. Figure 1 depicts the original Xception model.

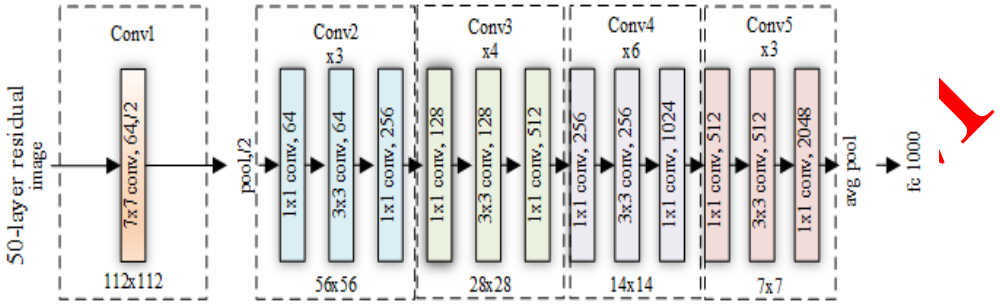

Figure 2. ResNet50 architecture

ResNet50 [40], developed by Microsoft Research in 2015, is a member of the ResNet family that achieved significant success in the ImageNet competition. The model takes video frames as input and utilizes preprocessing techniques such as normalization and data augmentation for deepfake detection. Through 50 layers, it extracts hierarchical features from the preprocessed frames. ResNet50 employs residual blocks to extract and learn features from images. The frames are resized to 256x256 pixels, and grayscale textural features are extended to optical flow fields. The model classifies the video footage as normal or abnormal after acquiring the feature vector. The ReLU activation function helps the model learn complex features. The model performs classification in the last layer by utilizing a fully linked layer and a softmax activation function. Figure 2 depicts the ResNet50 architecture.

Xception and ResNet50 architectures are effectively used in detecting fake images and videos. During training, both models are trained on extensive datasets consisting of real and deepfake video frames, enabling them to distinguish between these two types of content. In the inference phase, each frame's likelihood of being a deepfake is calculated, and these probabilities are combined to determine whether the video as a whole is a deepfake. By leveraging the power of deep neural networks, these architectures provide high accuracy in detecting fake videos.

3.4. Machine Learning For Classfication

The detection of deepfake videos is a matter of significant importance for digital security. Due to the widespread occurrence of these fake videos, numerous advanced technological methods have been developed to detect and distinguish them from the genuine videos. For instance, analyzing video metadata can provide valuable insights about the production process of these videos and reveal inconsistencies that indicate forgery [41]. Additionally, the process of examining video frames or optical flow for irregularities can help identify subtle characteristics that may go unnoticed by the human eye, hence assisting in the detection of fraudulent content.

SVM is a prominent ML method employed for the detection of deepfake videos [42]. It is a powerful and flexible ML algorithm extensively used in various classification and regression tasks. The SVM algorithm stands out from other classifiers because of its effectiveness in high-dimensional datasets. SVM employ hyperplanes to separate the data and to determine decision boundaries, thereby addressing the issue of high-dimensional data. Furthermore, SVM aim to reduce the empirical error while preserving the complexity of the mapping function.

The motivation for using the SVM to classify deepfake videos lies in its robustness and ability to handle overfitting issues. These characteristics enhance SVM's generalization ability, allowing it to achieve high performance on new, unseen data samples. SVM's ability to generalize predictions increases its performance in detecting deepfake videos, making it a reliable tool in digital security [43].

KNN is a straightforward and efficient algorithm used in supervised learning. It classifies new samples by comparing them to the closest training instances in the feature space. Upon the introduction of the test data, the algorithm initially identifies the nearest neighbors of this data in the feature space. Subsequently, the test example is assigned to the class that has the highest number of members among these neighbors [44].

MLP is a highly effective algorithm for performing classification tasks. This model process key features in the data through layers to make optimal predictions [45].

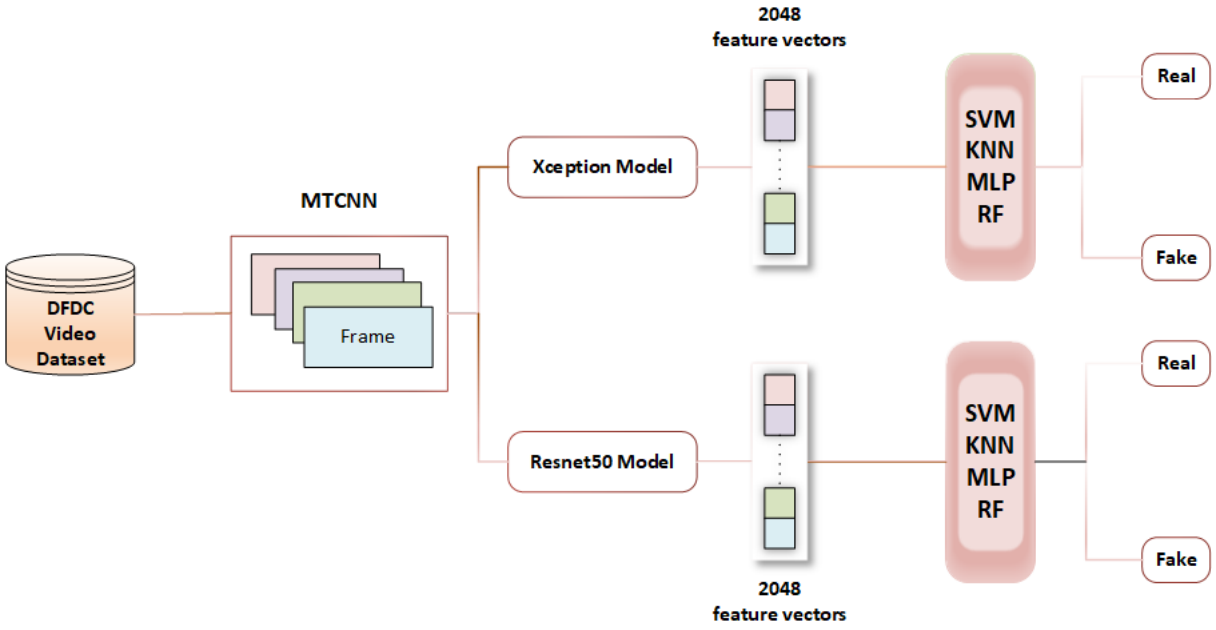

Figure 3. Flow diagram of our study

A MLP structure consists of three layers: input layer, hidden layer and output layer. The input layer receives the data, while the hidden layers perform computations to extract key features from the data. Finally, the output layer utilizes this processed data to produce the final predictions.

RF is a decision tree-based algorithm that generates multiple classification decision trees on different subsets of the dataset and combines them to make the final prediction by averaging. This method helps reduce overfitting issues. Furthermore, the RF algorithm is employed to determine the importance of features by adding the gain of each feature and scaling by the number of samples that pass through the node [45].

Figure 3 illustrates the flow diagram of our study. The DFDC dataset, which provides a wide range of data for deepfake vide detection, was utilized. The MTCNN method was employed to extract frames from selected real and fake videos. The Xception and ResNet50 were used to extract features from the obtained frames. Features were extracted from each frame, resulting in a 2048-dimensional feature vector for each frame. These extracted frames were saved, and CSV file was created according to the metadata file published by DFDC to ensure correct labeling and enhance the accuracy of the frames. Four frequently employed ML algorithms in deepfake video detection were utilized to classify the obtained feature vectors.

4. RESULTS

In this study, the DFDC dataset was used to test the proposed model. The dataset comprises a total of 5214 videos. To carry out the training and testing phases of our study, frames were extracted from the video set. The MTCNN method was utilized to extract the face images from the videos. Figure 4 shows frame examples obtained from the video. The labels for the 4765 extracted images were generated using the metadata file of the dataset.

4.1. Implementation Details

The Xception and ResNet50 models are highly successful in tasks such as image classification and are capable of extracting highly effective features that are particularly useful for the detection of deepfake videos.

Figure 4. Video frames, Image_a and Image_b are fake frames, Image_c and Image_d are real frames

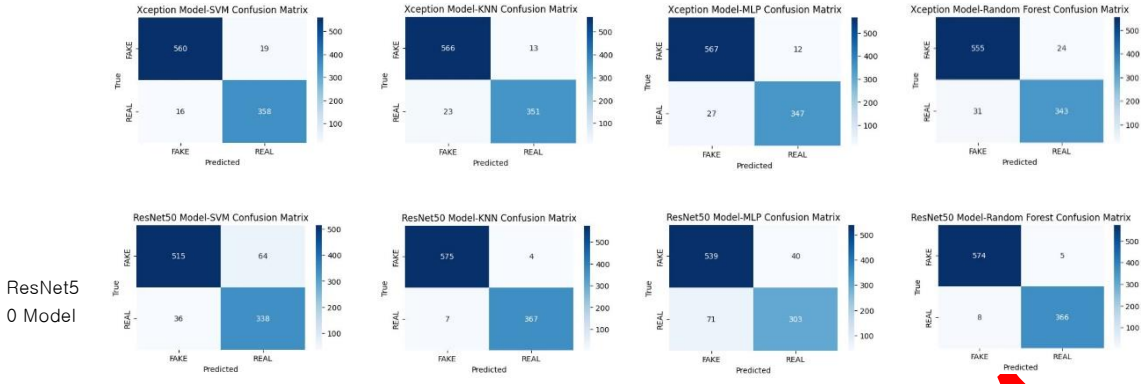

Figure 5. Confusion matrix outputs of classification methods

The objective of this work is to minimize computational cost and model complexity while maintaining high accuracy by exclusively utilizing these models during the feature selection phase. By employing various ML algorithms in the classification phase, we can compare the performance of various models and select the most suitable approach.

Features were extracted and saved from each of the 4765 images. These extracted features, along with the real-fake labels, were matched and used in the classification phase. The obtained face images were split into training and testing datasets with an 80%-20% ratio. The number of frames used in the training and testing phases, along with the total number of frames, is presented in Table 2.

Table 2. Number of frame

Frames/cropped faces in	Real	Fake
Total	3335	1430
Train Set	2756	1056
Test Set	579	374

Utilizing ML for classification purposes yields improvements in both performance and processing time reduction. The classification techniques employed in this study include SVM, KNN, MLP, and RF. Figure 5 represents the confusion matrices for the test results of these algorithms.

4.2. Evaluation Metrics

From the confusion matrices, the True Negative (TN), True Positive (TP), False Negative (FN) and False Positive (FP) values were obtained. Using these values, the accuracy and performance metrics, including F1-score, precision, and recall, were calculated. The equations used to calculate the afore-mentioned metrics are given below. Additionally, the accuracy rates of the proposed hybrid models are presented in Figure 5.

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

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (2)$$

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

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

Figure 6 illustrates the accuracy values of four different classification algorithms (SVM, KNN, MLP, RF) employed for deepfake image classification using Xception and ResNet50 models. In the classifications made with the Xception model, the SVM and KNN algorithms achieved the highest accuracy rate of 96.0%, followed by the MLP method with 95.0%, and the RF algorithm with 94.0%. When using the ResNet50 model, the KNN and RF algorithms provide best performance with an accuracy rate of 98.0%. In contrast, the SVM and MLP algorithms achieved lower accuracy rates of 89.0% and 88.0%, respectively.

These results indicate that KNN and RF algorithms with ResNet50 models achieve the highest accuracy in deepfake image classification. However, the total performance can differ depending on the combination of the model and the algorithm.

Accuracy represents the ratio of correctly classified examples to the total number of examples, and it is used as a general indicator of the model performance. Nevertheless, in imbalanced datasets, accuracy alone may not be sufficient. Therefore, other metrics such as F1-Score, Precision and Recall should also be considered when evaluating the proposed model. Especially in critical areas like deepfake detection, proposed models need to perform consistently and balanced. By considering these metrics alongside accuracy, a more accurate assessment of the model's efficacy in real-world application can be obtained.

Table 3 presents the performance metrics for real-fake classification for each ML method. F1 score is the value calculated taking into account both precision and recall and helps evaluate model performance. It measures the balance between the model's correct predictions and incorrect predictions. Precision refers to the ratio of correctly predicted positive samples to the total samples predicted as positive. Recall refers to the ratio of correctly predicted positive samples to the total true positive samples.

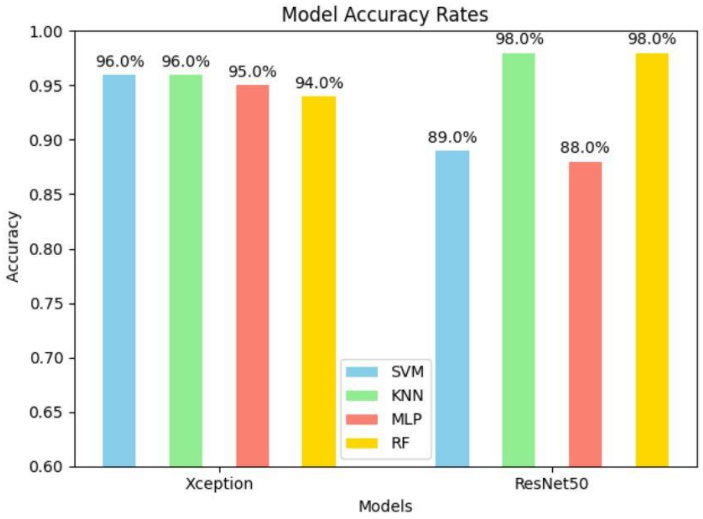

Figure 6. Accuracy rates of classification methods

Table 3. Metric rates of classification methods

Classification Methods		Xception Model			ResNet50 Model		
		F1-score	Precision	Recall	F1-score	Precision	Recall
SVM	Real (1)	0.95	0.95	0.96	0.87	0.84	0.90
	Fake (0)	0.97	0.97	0.97	0.91	0.93	0.89
KNN	Real (1)	0.95	0.96	0.94	0.99	0.99	0.98
	Fake (0)	0.97	0.96	0.98	0.99	0.99	0.99
MLP	Real (1)	0.95	0.97	0.93	0.85	0.88	0.81
	Fake (0)	0.97	0.95	0.98	0.91	0.88	0.93
RF	Real (1)	0.93	0.93	0.92	0.98	0.99	0.98
	Fake (0)	0.95	0.95	0.96	0.99	0.99	0.99

Based on the results of deepfake video classification utilizing Xception and ResNet50 models, the Xception model consistently exhibits superior performance on fake images. In evaluations using SVM, KNN, MLP, and RF algorithms, the Xception model offers higher F1-Score, precision, and recall values for fake images, while also performing well on real images. The ResNet50, on the other hand, achieved high performance with KNN and RF algorithms but demonstrated lower performance with SVM and MLP algorithms. Specifically, the KNN and RF algorithms demonstrated superior performance in both models. These findings indicate that the effectiveness of the model and algorithm employed in deepfake detection can differ, and different combinations may yield the best outcomes.

After in-depth examinations on the studies in the literature, it is observed that in addition to the fundamental Accuracy (ACC) metric, the AUC value is also used to evaluate the proposed models. AUC is considered as an important metric when evaluating the overall performance of a model in ML tasks for image classification. It is employed to thoroughly assess the performance of a model due to its independence from the threshold, its reliability even when the dataset is imbalanced, and its ability to summarize the classification capability of the model. Table 4 represents the obtained AUC values of the proposed hybrid models.

Table 4. AUC rates of classification methods

Classification Methods	Xception	ResNet50
SVM	%99.07	%94.74
KNN	%99.17	%99.47
MLP	%99.25	%95.45
RF	%98.76	%99.65

AUC values in Table 4 shows that the KNN and MLP algorithms demonstrated higher performance on classifications performed with Xception model. On the other hand, in classifications performed with ResNet50 model, the KNN and RF algorithms achieved better results. The findings indicate that combining the KNN and RF algorithms with ResNet50 model yields the most effectiveness in deepfake video detection.

4.3. Results and Discussion

This section presents a comparative analysis of our best model with existing studies in the literature. Table 5 provides the studies and their performance metrics, accompanied by detailed information on each.

A comprehensive analysis of the studies reveals significant differences in frame extraction, face detection, model training, classification method, and the datasets used. Feature extraction is identified as a pivotal stage in the process of fake video detection, as it plays a crucial role in model training.

Table 5. Literature review table

Ref.	Dataset	Frame Extraction	Model	Classification	Performance Metrics
[9]	Celeb-DF	Not mentioned	NA-VGG	X	AUC %85.7
[12]	DFD DF-TIMIT FF++ WildDeepFake	MTCNN	Xception	CNN	ACC %98.3
[18]	DFDC	DL libraries	CNN-Vision Transformer	Softmax	ACC %91.5
[19]	FF++ Celeb-DF UADFV	dlib	DeepFakeUCL	SVM	AUC %98.9
[22]	Celeb-DF DFDC DeeperForensics-1.0	dlib	Xception	Discriminator	ACC %96.0
[23]	SR-DF Celeb-DF	dlib	Multi-scale Transformer	ForgeryNet	AUC %95.5
[24]	Celeb-DF	Pyhton cv2	ResNet	LSTM	AUC %88.8
[25]	FF++ Celeb-DF DFD DeeperForensics-1.0	MTCNN	Inceptionv3	Siyam ağı	ACC %99.7
[27]	Celeb-DF DFDC	MTCNN	Vision Transformer EfficientNet	CNN	AUC %99.3
[28]	FF++ Celeb-DF DFDC	MTCNN	STA Masked Relation Learning	Temporal Convolution Network	ACC %91.81
[29]	FF++ DFDC Celeb-DF World Leaders dataset Cross dataset	MTCNN	GraphNet	Softmax	ACC %97.16
[30]	FF++ DFDC DFD	Not mentioned	PRNU spot analyzed	X	AUC %93.7
[31]	FF++ DFDC	dlib	Xception ResNet Inceptionv3	Softmax	ACC %98.5
[32]	FF++ Celeb-DF DFDC	MTCNN	Xception	MLP	ACC %98.00
[35]	DFDC	BlazeFace	EfficientNet	Softmax	AUC %91.8
Our Study	**DFDC**	**MTCNN**	**ResNet50**	**RF**	**ACC %98.0** **AUC %99.65**

Upon examining the ACC and AUC values of the eight hybrid models proposed in the study, it was observed that the best performance was achieved with the ResNet50+RF hybrid model. According to Table 5, considering the performance metrics of the studies using the same dataset, it is evident that our proposed hybrid model exhibits high performance. Furthermore, comparisons with the studies using different datasets showed that our model achieved consistently either outperformed or produced results that were close with those of the prior studies. Also, irrespective of the preferred dataset, our proposed model attains competitive results with existing literature and exhibits superior performance in the detection of deepfake videos.

5. CONCLUSION

Recently, deepfake technology has advanced significantly, enabling the creation of highly realistic fake audio, video, and image content. These materials present a considerable risk, as they have the potential to facilitate impersonation, the dissemination of misinformation, and a range of national security concerns that could compromise identity verification. This study proposes a hybrid model for the identification of fake content generated through the use of deepfake technology in video datasets. The objective is to mitigate the risks associated with the creation of content using deepfake technology. The hybrid model has been developed through the integration of a variety of ML methods, with a particular focus on DL algorithms.

The initial stage was the extraction of frames using the MTCNN method on the DFDC dataset. This was followed by the extraction of features using the Xception and ResNet50 models. For the classification stage, the SVM, KNN, MLP, and RF methods were employed. Upon testing the models obtained by combining these methods, it was found that the highest accuracy value was 98.0%, achieved through the hybrid combination of ResNet50 and RF algorithms. Furthermore, the same hybrid model achieved the highest AUC value, 99.65%.

The results indicate that the hybrid model proposed in this study outperforms existing models in the literature and is capable of accurately detecting deepfake content. The proposed model provides an effective solution to the challenges and threats posed by deepfake technology and represents a significant advancement in deepfake video detection. It is crucial to prioritize the development of real-time and low-latency systems for detecting deepfake content in order to enhance practical applications in future studies. The exploration of optimization approaches and hardware acceleration technologies has the potential to improve the computational efficiency of the model.

DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Aynur KOÇAK: Writing-Original Draft, Review, Editing, Methodology, Conceptualization

Mustafa ALKAN: Supervision, Methodology, Review, Editing, Validation

Muhammed Süleyman ARIKAN: Software, Methodology, Writing-Original Draft, Editing

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] M. Nawaz, Z. Mehmood, M. Bilal, A. M. Munshi, M. Rashid, R. M. Yousaf, et al., "Single and multiple regions duplication detections in digital images with applications in image forensic", ***Journal of Intelligent & Fuzzy Systems***, vol. 40, pp. 10351-10371, (2021).

[2] T. Nazir, A. Irtaza, A. Javed, H. Malik, A. Mehmood, and M. Nawaz, "Digital image forensic analysis using hybrid features", in ***2021 International Conference on Artificial Intelligence (ICAI)***, pp. 33-36, (2021).

[3] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security", ***California Law Review***, vol. 107, p. 1753, (2019).

[4] FaceApp. Available: https://www. faceapp.com/ (12.06.2024).

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

[6] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, et al., "The deepfake detection challenge (dfdc) dataset", ***arXiv preprint*** arXiv:2006.07397, (2020).

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial net", ***Advances in Neural İnformation Processing Systems***, vol. 27, (2014).

[8] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection", ***Iet Biometrics***, vol. 10, pp. 607-624, (2021).

[9] X. Chang, J. Wu, T. Yang, and G. Feng, "Deepfake face image detection based on improved VGG convolutional neural network", ***39th Chinese Control Conference***, pp. 7252-7256, (2020).

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", ***arXiv preprint*** arXiv:1409.1556, (2014).

[11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "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, (2020).

[12] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection", ***in Proceedings of the 28th ACM International Conference on Multimedia***, pp. 2382-2390, (2020).

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks", ***IEEE Signal Processing Letters***, vol. 23, pp. 1499-1503, (2016).

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", ***in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition***, pp. 1251-1258, (2017).

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., "Imagenet large scale visual recognition challenge", ***International Journal of Computer Vision***, vol. 115, pp. 211-252, (2015).

[16] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection", ***arXiv preprint*** arXiv:1812.08685, (2018).

[17] A. G. Nicholas Dufour, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler, "Deepfakes detection dataset by google & jigsaw", (2019).

[18] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer", ***arXiv preprint*** arXiv:2102.11126, (2021).

[19] S. Fung, X. Lu, C. Zhang, and C.-T. Li, "Deepfakeucl: Deepfake detection via unsupervised contrastive learning", ***in 2021 International Joint Conference on Neural Networks (IJCNN)***, pp. 1-8, (2021).

[20] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images", ***in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition***, pp. 1-11, (2019).

[21] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses", ***in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)***, pp. 8261-8265, (2019).

[22] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection", ***in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition***, pp. 18710-18719, (2022).

[23] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, et al., "M2tr: Multi-modal multi-scale transformers for deepfake detection", ***in Proceedings of the 2022 International Conference on Multimedia Retrieval***, pp. 615-623, (2022).

[24] V. V. V. N. S. Vamsi, S. S. Shet, S. S. M. Reddy, S. S. Rose, S. R. Shetty, S. Sathvika, et al., "Deepfake detection in digital media forensics", ***Global Transitions Proceedings***, vol. 3, pp. 74-79, (2022).

[25] S. Kingra, N. Aggarwal, and N. Kaur, "SiamNet: exploiting source camera noise discrepancies using Siamese network for Deepfake detection", ***Information Sciences***, vol. 645, p. 119341, (2023).

[26] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection". ***in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition***, pp. 2889-2898, (2022).

[27] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "Deepfake detection algorithm based on improved vision transformer", ***Applied Intelligence***, vol. 53, pp. 7512-7527, (2023).

[28] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for deepfake detection", ***IEEE Transactions on Information Forensics and Security***, vol. 18, pp. 1696-1708, (2023).

[29] F. Khalid, A. Javed, H. Ilyas, and A. Irtaza, "DFGNN: An interpretable and generalized graph neural network for deepfakes detection", ***Expert Systems with Applications***, vol. 222, p. 119843, (2023).

[30] L. Zhang, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Unsupervised learning-based framework for deepfake video detection", ***IEEE Transactions on Multimedia***, 25, pp. 4785-4799, (2022).

[31] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, "A machine learning based approach for deepfake detection in social media through key video frame extraction", ***SN Computer Science***, vol. 2, pp. 98, (2021).

[32] S. Mohiuddin, K. H. Sheikh, S. Malakar, J. D. Velásquez, and R. Sarkar, "A hierarchical feature selection strategy for deepfake video detection", ***Neural Computing and Applications***, vol. 35, pp. 9363-9380, (2023).

[33] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer", ***Advances in Engineering Software***, vol. 69, pp. 46-61, (2014).

[34] B. Doğan and T. Ölmez, "A new metaheuristic for numerical function optimization: Vortex Search algorithm", ***Information Sciences***, vol. 293, pp. 125-145, (2015).

[35] S. Korkmaz and M. Alkan, "Derin öğrenme algoritmalarını kullanarak deepfake video tespiti", ***Politeknik Dergisi***, vol. 26, pp. 855-862, (2023).

[36] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts", ***Ieee Access***, vol. 7, pp. 41596-41606, (2019).

[37] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment", ***in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition***, pp. 7184-7193, (2019).

[38] D. Huang and F. De La Torre, "Facial action transfer with personalized bilinear regression", ***in Computer Vision–ECCV 2012: 12th European Conference on Computer Vision***, pp. 144-158, (2012).

[39] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models", ***in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition***, pp. 9459-9468, (2019).

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", ***in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition***, pp. 770-778, (2016).

[41] J. Pu, N. Mangaokar, L. Kelly, P. Bhattacharya, K. Sundaram, M. Javed, et al., "Deepfake videos in the wild: Analysis and detection", ***in Proceedings of the Web Conference 2021***, pp. 981-992, (2021).

[42] F. F. Kharbat, T. Elamsy, A. Mahmoud, and R. Abdullah, "Image feature detectors for deepfake video detection", ***in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)***, pp. 1-4, (2019).

[43] M. Masood, M. Nawaz, A. Javed, T. Nazir, A. Mehmood, and R. Mahum, "Classification of Deepfake videos using pre-trained convolutional neural networks", ***in 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)***, pp. 1-6, (2021).

[44] N. Chakravarty and M. Dua, "A lightweight feature extraction technique for deepfake audio detection" ***Multimedia Tools and Applications***, pp. 1-25, (2024).

[45] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, et al., "Deepfake audio detection via MFCC features using machine learning", **IEEE Access**, vol. 10, pp. 134018-134028, (2022).