BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 13, No. 3, September 2025

Research Article

Deepfake Detection for Digital Image Security using Deep Learning Methods

Huseyin Alperen Dagdogen, Resul Das, Ibrahim Turkoglu

***Abstract*—With the rapid increase and proliferation of digital technologies, manipulated content is also increasing in parallel. With the widespread use of deepfake technology, the detection of manipulated content and deceptive content reduces the risks of manipulated data. This situation leads to serious security consequences at social, political, and personal levels with the creation of fake news, misleading videos, and audio recordings. This technology can also cause serious problems such as malicious use and violation of privacy. Therefore, it is vital to develop preventive measures such as deepfake detection and to use this technology correctly and ethically. The detection of images created using deepfake techniques aims to detect manipulations in media files such as video and audio using artificial intelligence and machine learning techniques. Deepfake detection is usually carried out using deep learning algorithms and models.**

In this study, a hybrid model consisting of transformer-based networks and Convolutional Neural Networks (CNNs) is used to classify fake and real images. When the results of the study were examined, it was seen that the hybrid model used gave more successful results compared to the literature. The applications were carried out on the Casia-WebFace dataset. According to the results obtained, the proposed artificial intelligence method plays an important role in the classification process of images produced using DeepFake techniques. 98.82% accuracy rate was achieved for the Casia-WebFace dataset. These results show that the proposed artificial intelligence model is effective and successful in predicting deepfake techniques.

***Index Terms*—DeepFake, Vision Transformer, Convolutional Neural Networks, Forgery, Deep Learning, Image Security.**

I. Introduction

WITH the advancement of technology, the digital world is becoming increasingly complex, and its boundaries are becoming blurred. The rise of deepfake technology further deepens this complexity. Deepfake appears as a phenomenon that emerges from the combination of manipulated images and sounds created using artificial intelligence and deep learning methods. However, the impacts of this new technology do not stop at the entertainment world; it also presents serious ethical, social, and security dimensions. DeepFake was initially used in the film industry to enhance effects, but over time it has become applicable in a wider area [1]. It appears in a wide range of areas, from works of art to political debates, from news to social media content. However, every aspect of this technology needs to be examined because the boundary between reality and fiction is becoming increasingly blurred with the manipulation of images and sounds. However, the spread of deepfake technology also brings with it some important concerns. In recent years, this technology, which has also been used in areas such as object tracking and detection [2] and determining attack types in large data sets, has led to important discussions in terms of reliability and ethical use [3]. The reliability of manipulated content in the fields of image processing and review and the potential for malicious use of this technology have become serious concerns. In this context, it has become very important to develop a conscious approach to the ethical use and security of deepfake technology. Additionally, the rapid advancement of technology has led to the emergence of a persistent security gap in today's world. These security vulnerabilities affect people in every aspect of life, whether social, political, or economic, and the consequences are severe. Particularly in recent years, the development of deepfake technology and the creation of fake images, videos, and audio files have made these issues a part of our lives. Initially, DeepFake technology was a technique aimed at replacing a targeted point on a video, image, or audio with another type [4]. Recently, this technology has advanced to the point where it can generate both fake and unique content in conjunction with generative AI. To prevent the complexity that has emerged with the proliferation of DeepFake technology, the detection of fake images is of extreme importance. This article aims to detect images produced by DeepFake technology using a new technology known as Transformer Networks, intending to prevent the mentioned issues. The results obtained from the experiments conducted in the study show that successful results can be achieved when the detection of fake images is done on a pixel basis, and additionally, the trained model can make a clear distinction between fake and real images.

A. Related Works

Deepfake detection studies include examining techniques developed to distinguish and detect fake content from reality. These studies focus on developing solutions to detect and block deepfake videos using technologies such as image processing, artificial intelligence, deep learning, and similar methods. Various studies in the literature have used deep learning models to determine the authenticity of videos and images in datasets.

Huseyin Alperen Dagdogen is with the Department of Software Engineering, Technology Faculty, Firat University, Elazig, 23119 Türkiye e-mail: hadagdogen@firat.edu.tr

Resul Das is with the Department of Software Engineering, Technology Faculty, Firat University, Elazig, 23119 Türkiye e-mail: rdas@firat.edu.tr

Ibrahim Turkoglu is with the Department of Software Engineering, Technology Faculty, Firat University, Elazig, 23119 Türkiye e-mail: iturkoglu@firat.edu.tr

Manuscript received Jan 21, 2025; accepted Feb 15, 2025.
DOI: 10.17694/bajece.1624564

TABLE I: Review of Studies in The Literature

References	Year	Model	Measured Metrics	Cons
[5]	2020	Xception - MobileNet	Accuracy	Does not use inter-frame correlations Performance results at low resolution not examined
[6]	2018	VGG Face	Accuracy AUROC	Parameters such as learning rate, batch size or dropout are not optimized Limited model comparisons have been made It used only a CNN-based approach
[7]	2020	MesoNet - CNN	Accuracy	The model used is a shallow model, not a deep one Has a limited number of parameters A lightweight and low computational power model
[8]	2019	SCNNet	Accuracy AUROC	Not compared with methods based on time series analysis The performance of the model at different threshold values has not been examined
[9]	2018	InceptionV3	Accuracy	Too little data Low resistance to adversarial attacks
[10]	2020	Deepfake Stack	Precision Recall F1 Score Accuracy AUROC	Too little data Low resistance to adversarial attacks
[11]	2022	CNN	Precision Recall F1 Score Accuracy	The model is based solely on CNN, no new generation model is used Insufficient dataset Lack of temporal analysis
[12]	2024	Inception-ResNet-V1	Accuracy	It is not stated how durable the model is in different conditions Imbalances in training data
[13]	2024	Xception	Accuracy ROC	Computational cost Lack of explainability Vulnerability to new deepfake techniques

In the study numbered [5], deep learning technologies called Xception and MobileNet were used to automatically detect DeepFake videos. Four datasets using four different and popular DeepFake technologies created by FaceForensics++ [14] were used as training and evaluation datasets. The results obtained in the study show variable accuracy between 91% and 98% depending on the level of development of deepfake technologies in the industry. In addition, the researchers developed a voting mechanism that can detect fake videos using a combination of four methods instead of just one method. However, since the application was performed on a high-resolution dataset such as FaceForensics++, generalization to low-resolution datasets is limited. When the studies in the literature are examined, it is seen that they are generally divided into four separate categories [15]. These techniques are generally deep learning-based techniques, classical machine learning-based methods, statistical techniques, and blockchain-based techniques. The studies also evaluated the performance of the detection capabilities of various methods according to different datasets and concluded that deep learning-based methods are more successful than other methods in detecting deepfakes [15]. Most of the studies in the literature generally focus on the production and detection of Dee. Most of the time, in these studies, algorithms such as Generative Adversial Networks (GANs) [16], autoencoders, Variational Autoencoders (VAEs) [17], Recurrent Neural Networks (RNNs) [18] and CNNs [19] are used for deep learning and generative artificial intelligence applications. Many of these algorithms mentioned in the literature are generally used for deepfake production. It is stated that algorithms such as CNNs, Long Short-Term Memory (LSTM), Capsule Networks, GANs, and Autoencoders are used for the detection of produced deepfake images. Especially in recent years, with the development of generative artificial intelligence, GANs models play an important role both in image generation and in the detection of fake images. In the study numbered [6], Deep Convolutional Generative Adversial Network (DC-GAN) model was selected to generate 64x64 images and Progressive Growing Generative Adversial Network (PG-GAN) models were selected for 256x256 and 1024x1024 images. The aim here is to help adapt the training data to various image sizes and qualities. In addition, VGG16 architecture was used for the classification process and the results were observed to be 80%. However, since the obtained success rate is low, it is not at the level to be used in forensic studies. Non-deep learning methods require lower computational power and data requirements. However, these methods cannot reach the accuracy and consistency rates of deep learning-based methods. [7] presents a deep learning-based deepfake detection method called MesoNet. MesoNet is a two-step process. A process known as feature extraction extracts facial features in the first stage. The second stage uses a classification process to classify these features as fake or real. The study also introduces a new activation function called Pish to improve the accuracy and consistency of MesoNet. With Pish, MesoNet achieves an accuracy rate of 86.89% on the FaceForensics++ dataset. Only the combination of Swish and Mish performs better (over 87%), followed by the baseline combination (ReLU + Leaky ReLU) with 86.76%. The paper "Detecting Deepfake-Forged Contents with Separable Convolutional Neural Network (SCNNet) and Image

Segmentation" by C.M.Yu, C.T.Chang, and Y.W.Ti published in 2019 presents a deep learning-based deepfake content detection method called SCNNet. The accuracy rates for acSCNNet are comparable to other deep learning-based techniques used to detect fake content. However, SCNNet can use less data and computational power due to its separability technique. Therefore, SCNNet is suitable for applications that do not need a lot of data or processing power.

Deepfake detections are not only performed on images but also on data such as video, audio, and signals. When examining the study numbered [20], it is observed that the published articles are generally categorized under the four mentioned headings, and the models have been tested on these types of data. Most studies have used similar datasets. Kothandaraman et al. aimed to classify fake and real faces using the Inception-ResNet-V1 hybrid model in their study [12]. The proposed model is pre-trained with the VGGFace2 dataset, thus performing face recognition with high accuracy rates. In the study, face extraction was performed from the dataset using the MTCNN (Multi-Task Cascaded Convolutional Networks) algorithm. In this way, the input data was made more regular and the model's learning rate contributed to the success. However, how the model would react to factors such as different lighting conditions, angles, and rotations was not tested and specified. Similarly, Joshi et al. used the Xception model to classify images and videos produced with deepfake technology in their study [13]. The Xception architecture, which uses depthwise separable convolutions rather than CNNs architectures, provides more effective results. A test accuracy rate of 93.01% was obtained in the study.

B. Main Contrubitions

The proposed hybrid model in this study provides significant contributions to the literature on the detection and classification of images produced by deepfake techniques. In particular, our approach provides new strategies that provide both higher prediction accuracy and improved interpretability. The main contributions of the study can be summarized as follows:

- The proposed model combines the strengths and advantages of both architectures by combining the traditional CNNs architecture and the new generation Vision Transformer (ViT) architecture. While CNNs capture features, ViT learns long-term dependencies. By hybridizing these two powerful architectures, challenging tasks can be detected with higher accuracy and generalization.
- The proposed hybrid model uses CNNs to optimize the high computational cost and create an efficient structure. Thus, it requires lower processing power than pure Transformer models and offers a stronger generalization ability than CNNs models alone.
- The proposed model achieved better accuracy than traditional CNN-based methods such as VGG16, SCNNet, and CNN-MesoNet. This demonstrates that transformer architecture can significantly improve the performance when combined with CNNs in image classification.

C. Paper Organization

This study consists of 4 main sections including the introduction. After the introduction, the second section presents the proposed approach and implementation steps, and the third section presents the findings and discussion. The last section concludes the paper. Figure 1 shows the relevant organization chart.

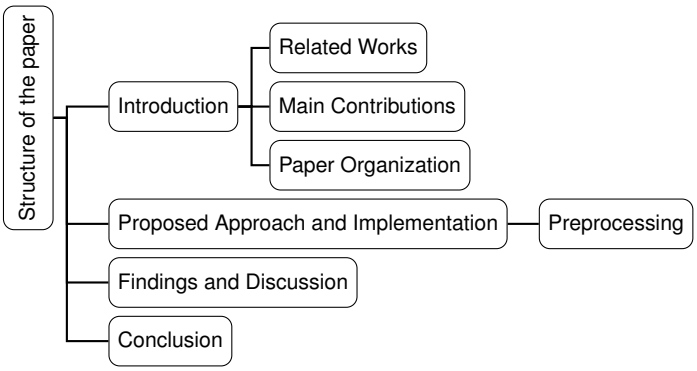


Fig. 1: Organization of the paper

II. Proposed Approach and Implementation

Convolutional neural networks have long been at the forefront of computer vision research. However, a noteworthy new development has surfaced recently in image processing. By modifying the transformer design, which has proven effective in Natural Language Processing (NLP), to analyze pixel-based data directly, ViT challenges conventional CNNs techniques. In essence, the ViT processes images using larger units known as "patches" rather than the traditional pixel-based method [21]. Using this method, the image is divided into patches, each of which is flattened before being fed into transformer blocks. ViT most distinctive characteristic is that, rather than using the convolutional and pooling layers seen in conventional CNNs, it uses transformer blocks that are rich with attention mechanisms. Figure 2 displays the model's architecture that was suggested for the investigation.

The study's suggested hybrid model combines the ViT model with convolutional neural networks. While ViT was used for the classification process, CNNs was used for feature extraction. Once the attributes of the input photos have been extracted, the ViT model has been utilized to classify the images and determine whether or not a certain image is fake.

The first step in the model is processing the images through several normalizing and convolutional layers. To reduce dimensionality, convolutional layers first extract feature maps from the images and subsequently identify patterns. After that, feature vectors are produced by flattening the feature maps. Activation, batch normalization, pooling, and convolutional layers make up the network architecture used for feature extraction. In the convolutional networks, the images begin at 512×512 dimensions and gradually shrink to 32×32 dimensions. The 512×512 pictures that pass through the convolutional layer first travel via the batch normalization layer before undergoing 2×2 maximum pooling operations.

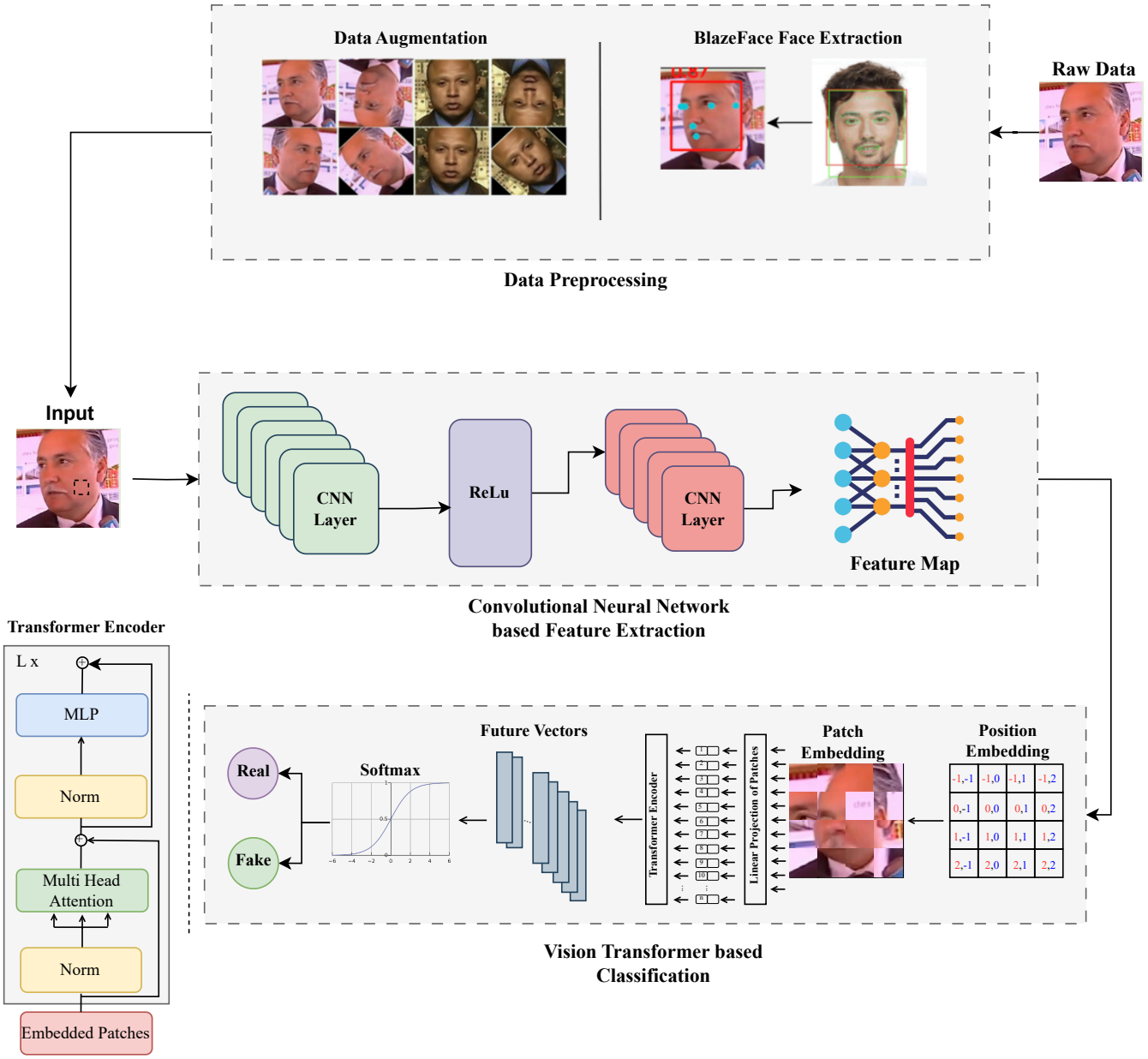


Fig. 2: The proposed hybrid model for deepfake detection

Features are captured sequentially by going through blocks of sizes 256×256, 128×128,64×64, and 32×32. For every size type, these operations are performed four times, using a stride value of 1.

A sequence of Transformer layers is applied to feature vectors. An attention mechanism pre-trained for better efficiency. The categorization task then made use of these relationships. The ViT network was used to train the captured features, and classification was carried out. Instead of processing the images' pixels directly, they use attention and self-attention mechanisms, which is different from standard CNN-based methods. Equation 1 displays the formula for the self-attention mechanism.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^{\intercal}}{\sqrt{d_k}}\right) \cdot \mathbf{V} \qquad (1)$$

As can be seen from the equation 1, Q stands for the query, K for the key, and V for the value matrices. The size of the key matrix is denoted by d_k. The self-attention mechanism's fundamental computational processes are illustrated in this formula. The six feedforward and attention layers that comprise the Transformer layers in the suggested model each have 1024-dimensional hidden units. Eight heads, each with 128 dimensions, are used in each attention layer. A linear layer has been incorporated into the photos after they were separated into patches of a particular size. The pictures have been split into 196 16×16 patches. Every patch has been transformed

Algorithm 1 A general interval ViT classification algorithm and CNNs feature extraction.

Function ViT()

function CNNFeatureExtract(data)
  for each image in data **do**
    Preprocess the image *e.g., resize, normalize*
    Pass the image through the CNN
    Extract features from the CNN's intermediate layer
    Store the features
  end for
end function
function VisionTransformerClassification(features, labels, validation_data, max_epochs, patience, initial_lr, factor, cooldown, min_lr)
  Initialize Vision Transformer model with initial_lr
  best_model ← Vision Transformer model
  best_loss ← ∞
  no_improvement ← 0
  current_lr ← initial_lr
  for epoch = 1 to max_epochs **do**
    Train the Vision Transformer using features and labels with current_lr
    Compute validation loss on validation_data
    if validation_loss ¡ best_loss **then**
      best_loss ← validation_loss
      best_model ← Vision Transformer model
      no_improvement ← 0
    else
      no_improvement ← no_improvement + 1
      if no_improvement ≥ patience **then**
        break *Early stopping*
      end if
    end if
    if no_improvement ≠ 0 and no_improvement % cooldown == 0 **then**
      current_lr ← current_lr * factor
      if current_lr ¡ min_lr **then**
        current_lr ← min_lr
      end if
      Update Vision Transformer model with current_lr
    end if
  end for
  return best_model
end function
Input: Training data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Extract features using CNNFeatureExtract($\mathcal{D}$)
Output: Extracted features $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$
Validation data $\mathcal{V} = \{(x_{v1}, y_{v1}), (x_{v2}, y_{v2}), \ldots, (x_{vm}, y_{vm})\}$
TrainedModel ← VisionTransformerClassification($\mathcal{F}$, $\mathcal{D}$, $\mathcal{V}$, max_epochs, patience, initial_lr, factor, cooldown, min_lr)
return TrainedModel

into 49 2048-size feature vectors.

Furthermore, the vectors used for position embedding have a size of 1024. Moreover, the model incorporates the positional interactions between patches by embedding positions for every patch. Classification is done using the output from the transformer layers. For classification purposes, the outputs from the transformer layers were routed to an output layer with softmax activation for two classes after passing via an intermediary layer of size 2048. The output was run through a classification head to identify the class to which the provided image belongs. Algorithm 1 displays the algorithm used in the pertinent investigation.

A. Preprocessing

In the study, data preprocessing was performed in two stages. These are face extraction and data augmentation processes. BlazeFace [22] model was used for face extraction. The BlazeFace Model developed by Google is lightweight and performs fast face detection. This model, which is optimized to run on devices with low hardware features such as embedded systems and mobile devices, performs face detection using deep learning techniques, especially since it is a low-dimensional model. BlazeFace model, which has a CNN-based architecture, extracts the features of the image with a lightweight backbone architecture and creates feature maps to define regions on the face. The extracted feature maps are subjected to a regression process in the next stage. In the regression process, whether the image is a face or not is classified from the feature maps. If the image is a face, key points are determined. The model enables fast processing for real-time use and offers high-accuracy face detection. A convolutional block of the BlazeFace model is shown in Figure 3.

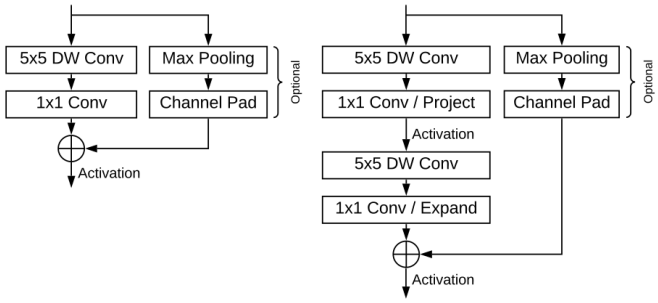


Fig. 3: BlazeBlock and Double BlazeBlock [22]

CNNs and deep learning were used to create face identification algorithm BlazeFace. A customized architecture that can identify objects quickly and accurately is the foundation of the operating principle [22]. The BlazeFace model's foundation, the Backbone Network, usually uses several backbone networks. These networks are employed for feature learning and data processing. These face detection-specific networks, such as MobileNetV1 [23] or MobileNetV2 [24], are often quick and lightweight models.

Traditional data augmentation methods have been applied in data augmentation procedures. Literature uses various data

augmentation approaches. These include methods such as data slicing, SMOTE, synthetic data generation, and augmentation. The data augmentation method used in this study helps to diversify the existing data set and provides a wider perspective to the model. Augmentation techniques in image processing include cropping, rotation, reflection, and brightness change. In this study, rotation was applied to the existing data set to perform data augmentation procedures. Different perspectives were obtained with these images rotated at certain angles. It is especially used in direction-independent object recognition operations. Thanks to the rotation of the data, the memorization of the training data of the model is prevented. Overfitting is prevented and the generalization ability of the model is improved with augmentation techniques such as rotation in small data sets. In this study, the images in the existing data set were rotated by 90 and 180 degrees, respectively, and recorded. An essential dataset for advancing facial recognition systems is the Casia-WebFace dataset [25]. Researchers utilize this dataset, a sizable database with various facial traits, to create and evaluate facial recognition systems. Facial photos of persons of various racial, age, and gender identities are included in Casia-WebFace. Additionally, it contains images captured in a variety of settings, including varying lighting, expressions, stances, and viewpoints. This kind is quite useful for assessing the performance of facial recognition systems in practical settings.

The Casia-WebFace collection contains 494,414 face images of 10,575 real identities collected from the web. These photographs cover a variety of facial features and were taken in a variety of environments. The effectiveness of face recognition systems in different environments is evaluated using this richness [26]. In the study, 80% of the dataset is used for training, 10% for validation, and 10% for testing. Figure 4 shows sample photographs from the Casia dataset and data augmentation examples.

Fig. 4: Casia WebFace Dataset Image Samples

III. Findings and Discussion

In the study, DeepFake classification was performed using CNNs and ViT architectures. The features extracted with CNNs were classified using ViT, and successful results were obtained. The proposed model has performed a faster and more accurate detection process. The differentiation of manipulated images has been achieved with a high success rate. In the preprocessing of the images, the BlazeFace face extraction

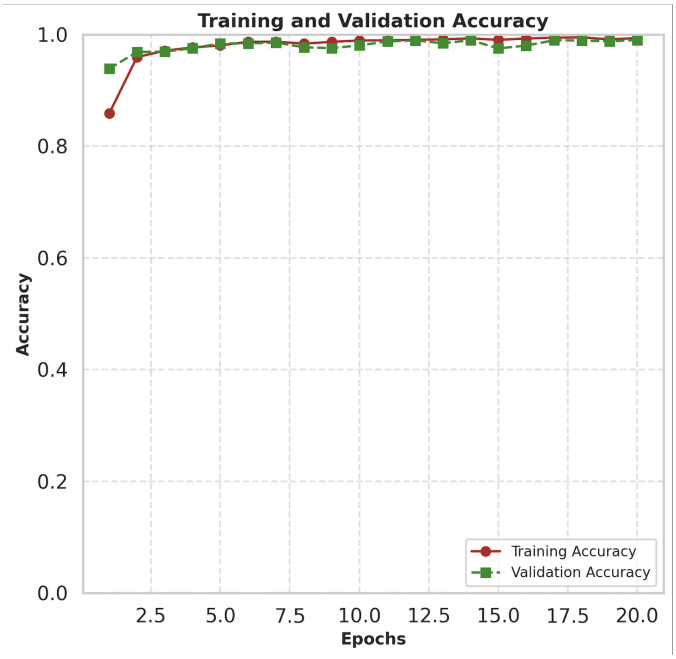


Fig. 5: Train accuracy graph

network was used, and existing data was augmented using data augmentation techniques. The data was trained using the EarlyStopping and ReduceLROnPlateau functions. The patience value for the EarlyStopping function was set to 5, and the patience value for the ReduceLROnPlateau function was set to 10. After all parameters were determined, the training was completed at 20 epochs, and the success rate for the CasiaWebFace dataset was observed to be 98.82%. The training accuracy graph of the study is shown in Figure 5, and the training loss graph is shown in Figure 6. Figure 5 shows that the graph shows that ViT can be an effective tool for DeepFake detection. The model can learn the training data quickly and the accuracy levels are quite high. The rapid increase in accuracy in the first 20 periods shows that the model learns the training data quickly and successfully learns the entire structure from simple features to more complex features. The proposed model uses the strengths of both models thanks to the hybridization of CNNs and ViT architectures. When CNNs, which have a successful structure in capturing features and relationships, are combined with the strong feature extraction ability of ViT capacity to learn global context and long-range connections, a high increase in the performance rate is achieved. The results obtained show that the study carried out by classifying the features extracted by CNNs with ViT can learn complex features that are important for distinguishing deepfakes. The accuracy fixed at approximately 98.82% shows that the model has completely learned the training data.

The parameters of the hybrid model proposed in the study are shown in Table II. In order for the attention mechanism to work more efficiently, the Feed Forward Network depth was determined as 6 and the number of neurons in the hidden layers was selected as 1024. The model, which consists of 8 attention heads, uses a multi-head attention mechanism. EarlyStopping was used to prevent overfitting and to reduce

TABLE II: Hyperparameter Description of Proposed Model

Model	Hyperparameters	Description	Value
ViT	Feed Forward Network	Artificial neural network without feedback or loop	6
	Dimensional Hidden Units	Number of neurons in the hidden layer	1024
	Number of Head	The number of topics an attention mechanism is divided into	8
	Patience Value for Early Stopping	To avoid overfitting and reduce unnecessary computational costs	5
	Patience Value for ReduceLROnPlateau	Allows the model to reach a better minimum	10
	Initial Learning Rate	The initial learning rate value used during training of the model	3e-4
	Learning factor	Used in learning rate schedulers in ReduceLROnPlateau	0.1
	Optimizer	is the algorithm that updates the weights to minimize the loss function	AdamW
CNNs	Input Layer	Layer representing the image data given to the model	512x512
	Kernel Size	Specifies the size of the filter used in the convolution layer	7x7
	Stride	Specifies how many pixels the kernel will shift on the input data	1
	Activation Function	Mathematical function that determines the output of neurons	ReLU
	Dropout	Regularization technique used to prevent overfitting	0.5

TABLE III: Proposed Model Accuracy Graph

Reference	Method (Classification)	DataSets	Accuracy	F1 Score	Recall	Precision
[5]	Xception - MobileNet	FaceForensics++	91%-98%	-	-	-
[7]	CNN- MesoNet	FaceForensics++	87%	-	-	-
[8]	SCNNET	FaceForensics++, Celeb-Df	95%	-	-	-
[6]	VGG16	CelebA-HQ	80%	-	-	-
[9]	Conv-LSTM	Hoha	97.01%	-	-	-
[10]	DeepFake Stack	FaceForensics++	99.65%	1.00	0.99	0.99
[11]	CNN	CasiaV2	92.23%	0.97	0.85	0.91
[12]	Inception-ResNet-V1	VGGFace2, DeepFake Dataset	97.5%	-	-	-
[13]	Xception	Kaggle deepfake_faces	93.01%	-	-	-
Proposed Model	CNN-ViT	Casia-WebFace	98.23%	0.97	0.98	0.97

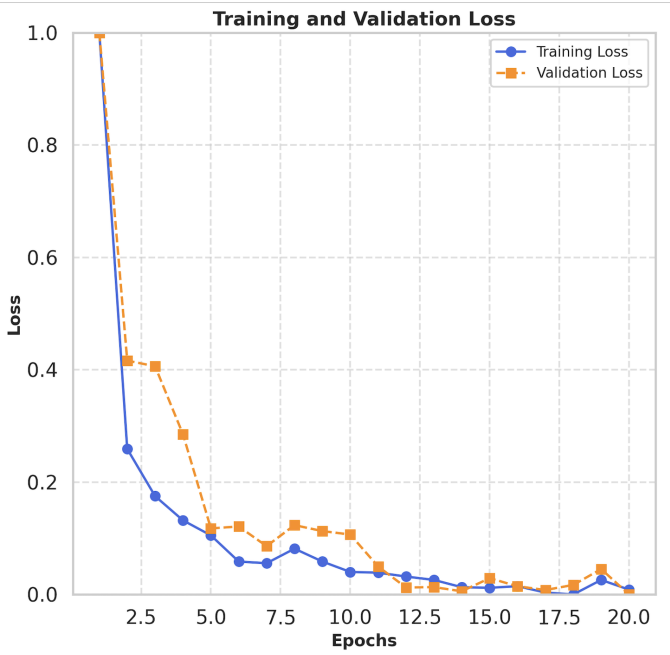


Fig. 6: Traning Loss Graph

unnecessary computational load. The EarlyStopping patience value was selected as 5. In addition, the ReduceLROnPlateau patience value, which is the learning rate planner, was selected as 10. In this way, the most optimal training hyperparameters of the model were selected with both EarlyStopping and ReduceLROnPlateau. Initially, the learning rate was selected as 3e-4 and the learning factor was determined as 0.1. In this way, the learning rate of the model was adjusted in a controlled manner during training.

As seen in the graph in Figure 6, the rapid decrease in the training loss in the first few epochs and the slowdown afterwards indicate that the learning process of the model is successful. As a result of 20 epochs of training with EarlyStopping, the training loss reached 0.2. The rapidly decreasing training loss shows that the model is learning fast. The slower decrease in the training loss after that indicates that the model starts to make smaller improvements. As seen in Figure 6, it is important to note that the training loss does not drop to 0, however, the low training loss obtained shows that the model is likely to perform well in general. Thus, when all results are taken into account, it shows that the proposed hybrid model for deepfake detection has a successful training process. The comparison of the results is shown in Table III. Although there are studies with higher success rates, the proposed model has an original structure thanks to a hybrid approach that demonstrates the strong feature extraction of the CNNs architecture and the classification power of the ViT architecture. Transformer models are particularly good at contextual learning on large datasets and therefore can perform better in image-based studies that contain more data. The proposed hybrid model combines traditional and new-generation deep learning techniques by hybridizing CNN and ViT architectures, providing a powerful representation of learning. While working on small datasets, the reduction of ViT's large data requirement and the preservation of its ability

to capture local features make the proposed approach unique and advantageous. One of the biggest advantages of the hybrid model is that it exhibits an effective approach to detecting time-varying inconsistencies. While CNNs and RNNs are successful in feature extraction and learning relationships between consecutive data points, Transformer-based models perform a more effective learning process since they capture spatial and temporal correlations better. In addition, the proposed hybrid model offers a fast and flexible learning process by providing scalability, parallelization, and efficient computation on large-scale image sets.

As shown in Table III, although CNN-based architectures show strong performance, their integration with new generation models such as GAN like ViT increases their performance rates. The proposed model CNNs-ViT provides a competitive accuracy rate compared to traditional CNNs models, indicating that ViT integration provides advantages. In model selection, high accuracy rates, used datasets and general model architecture should be taken into consideration.

IV. Conclusion

The analysis of the findings of the study shows that the hybrid classification model based on ViT and CNNs can work well to detect images generated by deepfake techniques. The model, which learns the training data quickly, performs the classification of fake images with high accuracy rates. In this article, local features of the images are extracted with CNNs, and maps of the extracted features are created. These maps, which are represented at high levels, are classified by ViT. The casia-WebFace dataset is used for training and testing in the study and 98.23% accuracy, 0.97 F1 Score, 0.98 Recall, 0.97 Precision values are obtained. Images generated by deepfake techniques, which have become a part of our lives today, create a major security gap. Many models have been proposed in the literature for the detection of fake images, but the model proposed in this study is among the models with high accuracy rates. In this way, the proposed model has great importance in terms of image security. As a result of the hybridization of CNNs and ViT architectures, the model explains the sensitivity. Especially with the increase in fake face images, these security vulnerabilities have become a bigger problem, and thus the model can become a new criterion for identifying these fake face images. The success of the proposed model is of great importance, especially when considering the security threats caused by deepfake technologies. With the increase in fake face images, security vulnerabilities such as identity theft, fake news, and fraud are increasing. In this context, the model's ability to detect fake images with high accuracy rates will contribute to ensuring security in the digital environment.

As a result, our hybrid model based on ViT and CNNs is considered as an important step towards increasing digital security by providing a reliable and effective method for detecting fake face images. The CNNs and ViT model opens the door to new research by showing that self-attention-based transformative models can be effective in deepfake detection applications. The proposed model can be tested with larger datasets in future studies to increase its generalization ability and to be integrated into real-time applications. The model, which is an important step for increasing digital security, will be tested with different datasets and trained with more realistic deepfake images, which will provide a more robust and comprehensive solution in terms of security.

Abbreviations

GANs Generative Adversial Networks
VAEs Variational Autoencoders
RNNs Recurrent Neural Networks
CNNs Convolutional Neural Networks
LSTM Long Short-Term Memory
DC-GAN Deep Convolutional Generative Adversial Network
PG-GAN Progressive Growing Generative Adversial Network
SCNNet Separable Convolutional Neural Network
NLP Natural Language Processing
ViT Vision Transformer

References

[1] F. Zengin, "Akilli makine çaği sinemasina giriş: Sinema sanatinda yapay zekâ teknolojilerinin kullanimi," *İletişim Çalışmaları Dergisi*, vol. 6, no. 2, pp. 151–177, 2020.

[2] R. Daş, B. Polat, and G. Tuna, "Derin Öğrenme ile resim ve videolarda nesnelerin tanınması ve takibi," *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 31, no. 2, p. 571–581, 2019.

[3] H. Ahmetoğlu and R. Daş, "Derin Öğrenme ile büyük veri kumelerinden saldırı türlerinin sınıflandırılması," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–9.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," *2018 10th Ieee International Workshop on Information Forensics and Security (Wifs)*, 2018. [Online]. Available: ⟨GotoISI⟩://WOS:000461290400003

[5] D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, "Deepfake detection through deep learning," in *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 2020, pp. 134–143. [Online]. Available: 10.1109/BDCAT50828.2020.00001

[6] N.-T. Do, I.-S. Na, and S.-H. Kim, "Forensics face detection from gans using convolutional neural network," *ISITC*, vol. 2018, pp. 376–379, 2018.

[7] P. Kawa and P. Syga, "A note on deepfake detection with low-resources," *CoRR*, vol. abs/2006.05183, 2020. [Online]. Available: https://arxiv.org/abs/2006.05183

[8] C. Yu, C. Chang, and Y. Ti, "Detecting deepfake-forged contents with separable convolutional neural network and image segmentation," *CoRR*, vol. abs/1912.12184, 2019. [Online]. Available: http://arxiv.org/abs/1912.12184

[9] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6. [Online]. Available: 10.1109/AVSS.2018.8639163

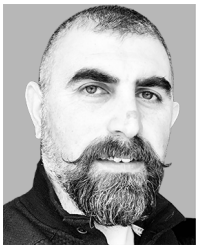
[10] M. S. Rana and A. H. Sung, "Deepfakestack: A deep ensemble-based learning technique for deepfake detection," in *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, 2020, pp. 70–75. [Online]. Available: 10.1109/CSCloud-EdgeCom49738.2020.00021

[11] S. S. Ali, I. I. Ganapathi, N.-S. Vu, S. D. Ali, N. Saxena, and N. Werghi, "Image forgery detection using deep learning by recompressing images," *Electronics*, vol. 11, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/3/403

[12] K. D, S. S. Narayanan, M. I. M, A. Yekopalli, and S. K. S, "Deep fake image classification engine using inception-resnet-v1 network," in *2024 International Conference on Computing and Data Science (ICCDS)*, 2024, pp. 1–5. [Online]. Available: 10.1109/ICCDS60734.2024.10560424

[13] P. Joshi and N. V, "Deep fake image detection using xception architecture," in *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)*, 2024, pp. 533–537. [Online]. Available: 10.1109/ICRTCST61793.2024.10578398
[14] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.
[15] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25 494–25 513, 2022. [Online]. Available: 10.1109/ACCESS.2022.3154404
[16] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.
[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[18] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.
[19] T. T. Nguyena, Q. Viet, H. Nguyenb, D. T. Nguyena, D. T. Nguyena, T. Huynh-Thec *et al.*, "Deep learning for deepfakes creation and detection: A survey," *SSRN Electron. J*, vol. 223, p. 103525, 2022.
[20] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1520, 2023.
[21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.
[22] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," *arXiv preprint arXiv:1907.05047*, 2019.
[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[25] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014. [Online]. Available: http://arxiv.org/abs/1411.7923
[26] J. Dong, W. Wang, and T. Tan, "CASIA image tampering detection evaluation database," in *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, Jul. 2013. [Online]. Available: https://doi.org/10.1109/chinasip.2013.6625374

Huseyin Alperen Dagdogen is a research assistant at Firat University's Software Engineering department. He graduated with B.Sc. and M.Sc. degrees from the Software Engineering Department of Fırat University in 2021 and 2024, respectively. After working for several private sector companies in 2021 and 2022, he was a research assistant at the Software Engineering Department of the Faculty of Engineering and Natural Sciences at Malatya Turgut Ozal University in 2023. After working at Malatya Turgut Ozal University for about 1 year, he started serving as a research assistant at the Software Engineering Department of the Faculty of Technology of Firat University in 2024. His current research areas include machine learning, deep learning, network systems, and graph systems.



Resul Das is a full professor in the Department of Software Engineering at the Faculty of Technology, Firat University. He received his B.Sc. and M.Sc. degrees in Computer Science from Firat University in 1999 and 2002, respectively, and completed his Ph.D. in Electrical and Electronics Engineering in 2008. Between 2000 and 2011, he served as a lecturer in the Department of Informatics and concurrently worked as a network and system administrator at the University's IT Center. Since 2002, he has been an instructor in the Cisco Networking Academy Program, delivering CCNA and CCNP courses. From September 2017 to June 2018, he conducted research as a visiting professor at the University of Alberta, Edmonton, Canada, under the TUBITAK-BIDEB 2219 Postdoctoral Research Fellowship program. He also held the position of Head of the Department of Software Engineering from March 2020 to April 2023.

Professor Das has served in editorial roles for several prestigious academic journals. He was an Associate Editor for IEEE Access and the Turkish Journal of Electrical Engineering and Computer Science. Currently, he is an Associate Editor for several Elsevier journals, including Internet of Things, Alexandria Engineering Journal, and Telematics and Informatics Reports, as well as the IEEE Open Journal of the Communications Society (OJ-COMS) and the International Journal of Grid and Utility Computing (Inderscience).

Globally recognized for his academic contributions, Prof. Das has been consistently listed among the top 2% of the "World's Most Influential Scientists," compiled by Stanford University researchers, for five consecutive years (2019-2024). His research interests include computer networks, cybersecurity, IoT and systems engineering, data science and visualization, and software quality assurance and testing.

Ibrahim Turkoglu graduated with B.S., M.S., and Ph.D. degrees from the Department of Electrical and Electronics Engineering at the Firat University in 1994, 1996, and 2002 respectively. He completed his doctoral and master's thesis on artificial intelligence and pattern recognition. In particular, he developed two artificial intelligence techniques in his doctoral thesis. He has published over two hundred papers in international conference proceedings and refereed journals and has been actively serving as a reviewer for international journals and conferences. He has also been involved in many national and international projects. His current research areas include artificial intelligence, pattern recognition, bioinformatics, signal/image processing, and embedded systems. He is a lecturer for many courses such as Artificial Intelligence, Embedded Systems, Pattern Recognition, and Deep Learning in the Software Engineering Department.