An Improved DeepFake Detection Approach with NASNetLarge CNN

İsmail İlhan
Vocational School of Technical Sciences
Adıyaman Univercity
Adıyaman, Turkey
0000-0002-5972-4295

Ekrem Balı
Department of Computer Engineering
Fırat University
Elazığ, Turkey
175260071@firat.edu.tr

Mehmet Karaköse
Department of Computer Engineering
Fırat University
Elazığ, Turkey
0000-0002-3276-3788

2022 International Conference on Data Analytics for Business and Industry (ICDABI) | 978-1-6654-9058-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDABI56818.2022.10041558

***Abstract*— Deep fake images are a new technology that has emerged with the development of computer vision and deep learning technologies in recent years. The development of these deep fake technologies has led to the production of many fake or manipulated products. Thus, the problem of detecting the deep fake has emerged and many methods have been developed to solve this problem. In this study, feature extraction and classification method on the dataset with NASNetLarge CNN deep learning model is proposed and a successful result is produced. In the proposed method, training and test datasets were created by removing facial regions from the video frames in the Celeb-DFv2 dataset. The architecture of the NASNetLarge model is explained and the success of the model is tested. According to the test results, an ACC value of 96.7% was obtained and compared with other methods. As a result, the study offers an easier model training with a smaller dataset than other methods and produces a competitive and successful result.**

Keywords—deepfake, manipulation, detection, NASNetLarge, image classification, video detection

I. INTRODUCTION

Today, image processing and deep learning technologies have made great progress. With the use of these two technologies together, many new methods have been created in the field of computer vision and deep learning techniques. One of these areas is synthetic image production . What is meant by synthetic image is deep fake, that is, deepfake images. Deep fake images are also produced using deep learning methods and especially GAN architecture [1], [2]. These produced images are used to create revenge porn and lies through politicians. This situation can have negative consequences for states, institutions and individuals. In order to prevent this, deep fake detection algorithms are implemented by using deep learning techniques. Although deep fake content generation is ahead of deep fake content detection, great success has been achieved in this regard [3]. Many different methods can be used for deep fake content detection. Classical machine learning and mostly deep learning methods are used. In deep learning methods, Convolutional Neural Networks (CNN) architectures are generally used. In addition, different methods such as Support-Vector Machine (SVM), which is a machine learning technique, that is, CNN with Support Vector Machine or RNN with CNN or LSTM with CNN can be used together. It is seen that deep fake detection processes can be achieved with different methods such as head position disorders, lip movements, feature extraction from facial regions – this method was used in this study – detection of biological signals, examination of sound signals in the detection of deep fake videos/images [4], [5]. In this study, the Celeb-DFv2[14] dataset, which is one of the most successful datasets, was used to detect deep fake videos. Face regions were extracted from the videos in the dataset with the DLib library. A new dataset consisting of images was prepared. Then, a deep fake detection model was created by retraining the NASNetLarge CNN model with the obtained dataset. The NASNetLarge model was created with automatic artificial intelligence to classify the CIFAR-10 dataset with the NAS (Neural Architecture Search) method, which is an AutoML (automatic machine learning) technique [6]. Then, the detection method created was tested with the ImageNet dataset and it was seen that the model was successful. The main contributions of this study are as follows:

(1) Proving the success of the NASNetLarge model, one of the CNN architectures, which is the most successful method in image extraction and classification among deep learning methods, in deep fake detection.

(2) To show that competitive results are obtained between the Celeb-DFv2 dataset, which is one of the most challenging datasets, and the studies using only facial feature extraction and classification methods in Deep fake detection methods, and to ensure a high success rate.

(3) To show that this problem can also be successfully solved with a smaller dataset.

II. RELATED WORK

In recent years, with the great development of computer vision and deep learning technologies, many methods have been developed for deep fake detection. Mostly CNN-based methods have been used to detect deep fake video or images.

In the studies of Şahin and Mustafa, feature extraction was made from video frames with the EfficientNet family, which uses CNN architecture. The resulting sequence is used as an input in a classifier network. In their study, they obtained a dataset by subtracting the facial regions from the images obtained from the video frames. In addition, certain data augmentation techniques were applied on it, then training of model parameters was disabled and training was conducted with EfficientNet. As a result of the experiments with the DFDC dataset, the highest accuracy value was obtained with EfficientNetB4 as 91.8%. As a result, this method shows that the EfficientNet model is a model that produces positive results for deepfake detection. The block diagram of the model is given in Fig.1 [7].

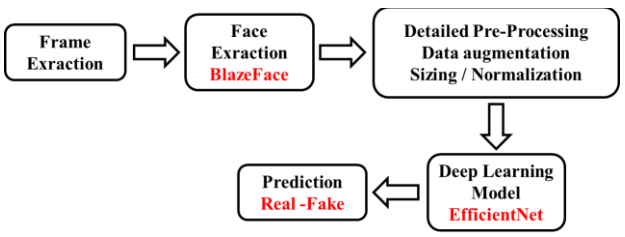

Fig. 1. Block Diagram of the EfficientNet Model [7]

In the study by Oliver et al., the traces left by the Generative Adversarial Network (GAN) engines during the creation of the deep fake are detected by analyzing the temporal frequencies. These unique fingerprints are called GAN Specific Frequencies (GSF). In order to capture these traces, the AC coefficients are calculated via the discrete cosine transform.

The analysis of these AC coefficients yields impressive classification results. In their studies, deep fake images were created with different methods by using real images in CelebA and FFHQ datasets, and a dataset was created from these data. These images have been replaced with different types of attacks such as mirroring, scaling, rotating, adding a random color, position and dimensional rectangle. High success has been achieved with this method .

Deep fake detection method in the work of Tianchen et al.; It is based on the hypothesis that source images can be preserved and extracted after going through a deep fake generation process. Pair-Wise Self-Consistency Learning (PCL) approach was used to train the CNN to detect deep fakes. Inconsistent Image Generator (I2G) is used to generate the rich annotated dataset required for PCL. This method generally uses inconsistencies of source features in fake images, unlike other methods that use artifacts. Conceptually, images can uniquely identify their source and carry spatially local information independent of the content. Therefore, a fake image contains different source images in different locations, while the real image is expected to be consistent in all locations. Deep fakes can be detected by extracting local source features and measuring the consistency of the image [8].

Aya et al. have designed methods by adding XGBoost classifier on proven CNN models called State-of-the-Art. The actual proposed method is shown as YOLO-InceptionResNetv2-XGBoost (YIX). With the YOLO face detector, the face region is extracted from the video frames. Feature extraction is done from these face images with InceptionResNetv2 and finally classification is done with XGBoost. In their studies, many different CNN models were tried to extract features and the best result was obtained with InceptionResNetv2. In their studies, a certain number of data was taken from the Celeb-DF and FF++ datasets and a common dataset was created. As a result, the 'accuracy' value was obtained as 90.0%. The block diagram of the model is given in Fig. 2 [9].

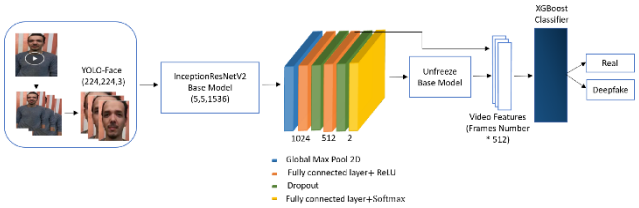

Fig. 2. Block Diagram of the YIX Model[9]

Dafeng et al. propose a deep fake detection model called DeepfakeNet, which consists of 20 network layers. It adopts ResNet's stacking idea and Inception's split-convert-merge ideas. In the data preprocessing module, the video dataset is processed first. Then the features of the facial images are extracted by CNN. After sufficient training and validation, the DeepfakeNet model is continuously improved to achieve better results. In order to obtain more accurate results, datasets are amplified by stretching, rotating, flipping and changing brightness methods, which are common methods of developing the dataset. The study uses some data from the FaceForensics++, Kaggle and TIMIT datasets, and 96.69% 'accuracy' has been obtained. The block diagram of the model is given in Fig. 3 [10].

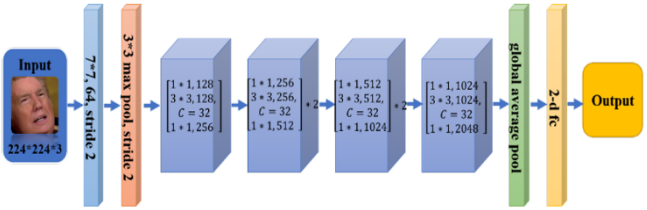

Fig. 3. Block Diagram of the DeepfakeNet Model [10]

Jin et al suggested examining biological signals for deep false detection. In their study, they extracted the biological signal difference between real and fake videos in three dimensions. They proposed a new detection method with multidimensional biological signals. Compared with other technologies, the proposed method only extracts fake video information and is not limited to a particular production method. Therefore, it is not affected by synthetic methods and has good adaptability. This article demonstrates that heart rate signals can be used to effectively distinguish between real and fake videos. For this reason, real and fake videos are classified by extracting and analyzing biological signals from videos. FF++, DFD, UADFV datasets were used for training the method [11].

CNN architectures, one of the deep learning techniques, are preferred in deep fake detection methods. In addition, architectures such as SVM [12], [13], RNN[14] and LSTM[15], [16] are also used in the methods. In addition, the most used datasets, frame numbers and creation methods for deepfake detection are shown in Table 1[17].

TABLE I. DATASETS, CREATION METHODS AND VIDEO-FRAME NUMBERS

Dataset	Real Video	Real Frame	Fake Video	Fake Frame	Method
UADFV	49	17.3B	49	17.3B	FakeAPP
DF-TIMIT	320	34B	320	34B	Faceswap-GAN
FF++	1000	509.9B	1000	509.9B	Deepfakes, Face2Face FaceSwap, NeuralTextures
DFD	363	315.4B	3068	2.2M	FF++'a Benzer
DFDC-Preview	1131	488.4B	4113	1.7M	Deepfake, GAN Nonlinear Methods
Celeb-DF	590	225.4B	5639	2.1M	Enhanced Deep Pseudo Synthesis Algorithm

The first datasets that have made major contributions to the growth and improvement of deep fraud detection technologies are UADFV[18] and DF-TIMIT[19]. FaceForensics++ [20] is created by real images downloaded from YouTube and forged using four different methods. Published by Google in collaboration with Jigsaw, the DeepFakeDetection [20] dataset (DFD) contains over 363 original videos from 28 different actors in 16 different scenes, as well as over 3000 manipulated videos using deep fakes. The DFDC[21] dataset consists of approximately 128 thousand videos and contains 10 million frames. An example dataset of

DFDC is given in the table. The Celeb-DF[22] dataset provides 590 real videos of mostly celebrities of different ages, ethnicities, and genders, and 5639 deep fake videos synthesized from them. In the Celeb-DFv2 dataset used in this study, there are 300 real videos taken from YouTube. Among the CNN-based feature extraction and classification methods, the proposed model achieves high success on the Celeb-DFv2 dataset by using less data. Accelerating the training process by reducing the dataset and developing high-successful results on the Celeb-DFv2 dataset are the contributions of the proposed model.

III. Proposed Method And Experiment Results

A. Proposed Method

Based on the researches in the literature, it is understood that deepfake detection with deep learning techniques is achieved with high success. As a result of these studies, the proposed method was determined as feature extraction and classification with CNN, and it was decided to use the NASNetLarge model, which has not been used for deep fake detection yet, and which is a proven method in image classification. In the proposed method, firstly, face regions will be extracted from the video frames in the Celeb-DFv2 dataset, and this dataset will be augmented with data augmentation techniques, then feature extraction and classification will be made over the dataset with NASNetLarge. The stages of the work performed are shown in Fig. 4. The main contribution of this study is that the NASNetLarge model is trained to detect deep fakes and is seen in the fourth step. The simple algorithm of the proposed model is as follows.

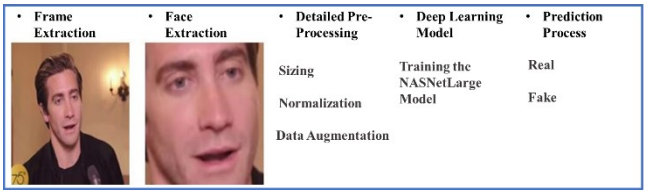

Fig. 4. Stages of the Model

Algoritm of proposed model

Step 1: Select input video from dataset
Step 2: Split frames from selected video and repeat step 3-6
Step 3: Apply face detection algoritm and exract face image from frame
Step 4: Apply image processing: sizing, rotation, normalization, filtering
Step 6: Train the NasNetLarge Model and update the detection deep learning weight or process the prediction
Step 7: Evaluate prediction rate

CNN architectures process the image in various layers. There is a convolution layer to detect features on the image. In this layer, masking with a filter is applied on the image and the features of the image are extracted according to the applied mask. For example, by masking, the edge regions of an image can be removed. After this layer, the pooling layer usually comes in CNN architectures. In this layer, a filter size is selected and this filter travels over the image, making it smaller by selecting the more important parts of the image. These two processes take place repeatedly and important and unimportant features are extracted to classify the image. The proposed method also uses the NASNetLarge model, which is a CNN model that works in this way.

B. NASNetLarge Architecture

Reinforcement learning search method is used in the proposed architecture and inspired by the Neural Architecture Search (NAS) framework [6]. Searching for the best convolutional architectures is reduced to searching for the best cell structure. The first type of convolutional cells is the Normal Cell and the second type is the Reduction Cell. These cell structures were created with the NAS model and accuracy values of 82.7% top-1 and 96.2% top-5 were obtained with the obtained NASNetLarge model. When creating the NASNetLarge model, the best convolution layer (or "cell") was searched in the CIFAR-10 dataset. These cells were then applied to the ImageNet dataset. Fig. 5 shows the CIFAR-10 and ImageNet cell architecture in the NASNetLarge model. Also, Fig. 6 shows the structure of the convolutional layer or cells.

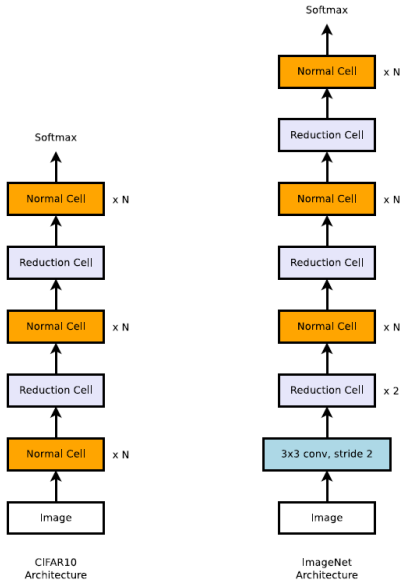

Fig. 5. Cell Architecture of the NASNetLarge Model [6]

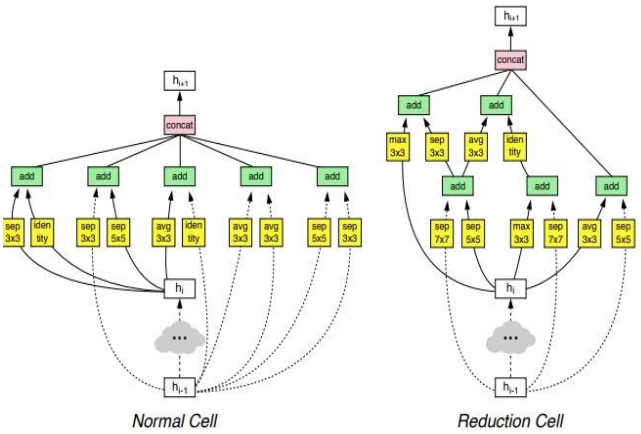

Fig. 6. Cell Structure of the NASNetLarge Model [6]

C. Layer Structure of the Proposed Method

Considering the successful results of this model in the ImageNet dataset, it was understood that it would be an effective CNN model for deep fake detection and was used for feature extraction in the proposed method. NASNetLarge parameters were completely retrained using the 'ImageNet' weights with the generated dataset. A flatten layer has been added to the end of the model to provide the outputs of this model as input to the classifier layer. Then, a dropout layer

was added to prevent overfitting. Finally, classification is done by adding a fully connected layer with one artificial neuron and its activation function 'sigmoid'. In addition, while compiling the model, 'binary_crossentropy', which is used in binary classification problems, is used as a loss function. 'Adam' was used as the optimizer. An optimized result was obtained by selecting the learning rate beta1 and beta2 parameters as '1e-5', 0.9, 0.999, respectively. Table II shows the model layers and parameters.

TABLE II. LAYER DETAILS OF THE PROPOSED MODEL

Layer (Type)	Output	Number of Parameters
NASNet (Functional)	(None, 11, 11, 4032)	84916818
Flatten_1 (Flatten)	(None, 487872)	0
Dropout_1 (Dropout)	(None, 487872)	0
Dense_1 (Dense)	(None, 1)	487873
Total Parameter : 85,404,691	Educable Parameter: 85,208,023	Ineducable Parameter: 196,668

D. Experiments and Results

1) Selecting the dataset

In the studies conducted by the researchers, the most used datasets were examined and more realistic fake videos were included in the Celeb-DFv2 dataset, which is the most difficult to distinguish with the human eye. In addition, in the studies examined, it was understood that the data set with the lowest success in deep fake detection methods with the CNN technique was Celeb-DFv2[5], [8]. For these reasons, the Celeb-DFv2 dataset was chosen and it was aimed to achieve high success. There are 5639 fake and 890 real videos in this dataset.

2) Preparation of the dataset

At this stage, firstly, each video was opened separately and facial regions were extracted using the DLib face detector. In this study, every frame in the videos was not processed, as it was desired to achieve success with a small data set. Approximately one face image was obtained from every second of the videos. Then, the data set was shrunk a little more by choosing from the obtained images. Under normal conditions, since there are many fake videos, the number of fake images is much higher. In the manual selection part, this situation was corrected, and a number of fake and real images were taken close to each other. Table III shows the number of images in the dataset.

TABLE III. DATASET DISTRIBUTION

	Fake	Real	Total
Train	8397	6726	15123
Validation	1200	962	2162
Test	1537	1560	3097

3) Training the model

The model proposed in this section is trained with the obtained dataset. Before the training started, many methods such as normalizing (reducing to 0-1 range), rotation, shifting, etc. from Image Augmentation methods were used on the data set and the data were randomly transformed. Then, an EarlyStopping function was set for the training of the model, and the training was started for 20 epochs by following the val_loss parameter. The training took a total of 12 epochs, the best result was obtained in the 9th epoch. Since there was no improvement in the 'val_loss' parameter after the 9th epoch, the training was stopped after the 12th epoch. Fig. 7 shows the 'accuracy' change of the model for 12 epochs, and Fig. 8 shows the 'loss' change.

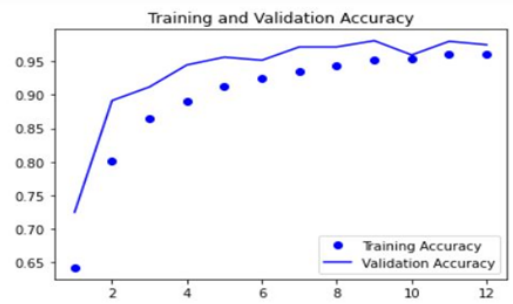

Fig. 7. Training and Validation Accuracy Graph of the Proposed Model

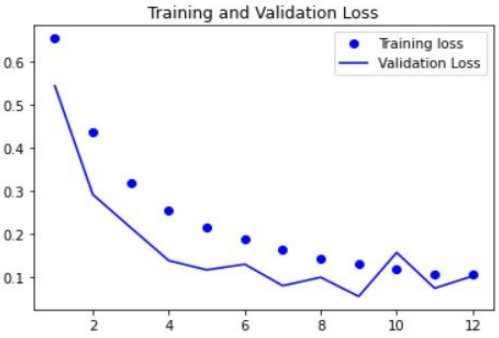

Fig. 8. Training and Validation Loss Graph of the Proposed Model

4) Model testing

Finally, the model was evaluated using test data, and the model success was obtained as 0.967 accuracy and 0.09 loss. When each of the test data entered the prediction process separately, high success was achieved, as can be seen in Table IV, and the prediction accuracy of the model was proven. As a result, approximately 93% of the predictions were successful. An example of the estimation process performed on the test data is shown in Fig 9. As it can be seen from the figure, although the Celeb-DFv2 dataset contains very realistic fake images, the classification is performed successfully.

TABLE IV. PREDICTION RESULTS OF TEST DATA

Number of Real Images with Correct Prediction	1519
Number of Real Images with Uncorrect Prediction	41
Number of Fake Images with Correct Prediction	1343
Number of Fake Images with Uncorrect Prediction	194

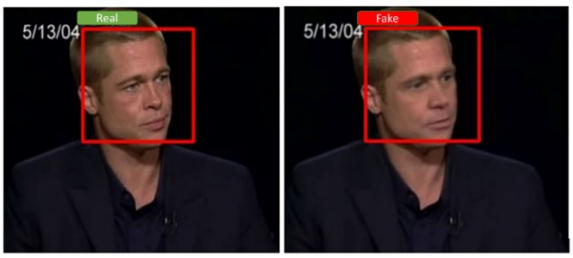

Fig. 9. Prediction Process with Trained Model

5) Comparison of the model with similar models

The proposed model clearly demonstrates its superiority over models using the Celeb-DF dataset. In addition to this, the reason why it lags behind some models is this. The Celeb-DF dataset is a more challenging dataset as it contains much better deep fake videos with less artifacts. When the datasets are examined visually, the Celeb-DF dataset is seen as the most incomprehensible dataset. In addition, from [5], [8], [22], [23] sources, it is seen that this dataset has lower success rates than others in deep fake detection. Table V shows the comparison.

TABLE V. COMPARISON OF THE PROPOSED MODEL

Method	Classifier	Dataset	ACC-AUC
EfficientNetB4[7]	CNN	DFDC	0.918
EfficientNetV2[24]	CNN	FF++ ve FFIW10K	0.97
YIX[9]	CNN - XGBoost	Celeb-DF ve FF++ (2848 Video)	0.906
DeepfakeUCL[8]	CNN	Celeb-DF	0.90
DeepfakeNet[10]	CNN	FF++ - Kaggle - TIMIT	0.96
NA-VGG[23]	CNN	Celeb-DF	0.85
NASNetLarge (Proposed Model)	**CNN**	**Celeb-DFv2**	**0.967**

IV. CONCLUSION

In this study, a model is proposed for the detection of deep fake videos/images, which is one of the important problems of today. This model is easy to train and has a high deepfake detection rate. The proposed model achieved the highest success in the most challenging Celeb-DFv2 dataset among the datasets created for deepfake detection by extracting features with the CNN architecture. In addition, while very large datasets are generally used to solve the deep fake detection problem, this study has shown that success can be achieved with a smaller dataset. The proposed method has achieved a competitive and high performance result when compared to methods other than feature extraction with CNN. As a result, when the model was tested, 0.967 'accuracy' and 0.09 'loss' were obtained, and then 93% correct classification was made with the estimation process on the test data. Among the CNN architectures and feature extraction and classification methods, the highest success was achieved with the Celeb-DF dataset. In future studies, it is aimed to increase this success by testing it with different classifiers and to produce an output that people can benefit from by transferring the project to the web environment.

REFERENCES

[1] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, 'Deepfake detection: a systematic literature review', *IEEE Access*, 2022.

[2] P. Yu, Z. Xia, J. Fei, and Y. Lu, 'A survey on deepfake video detection', *IET Biom*, vol. 10, no. 6, pp. 607–624, 2021.

[3] S. Ataş, İ. İlhan, and M. Karaköse, 'An Efficient Deepfake Video Detection Approach with Combination of EfficientNet and Xception Models Using Deep Learning', in *2022 26th International Conference on Information Technology (IT)*, 2022, pp. 1–4.

[4] N. Guhagarkar, S. Desai, S. Vaishyampayan, and A. Save, 'Deepfake Detection Techniques: A Review', 2021.

[5] İ. İlhan and M. Karköse, 'A Comparison Study for the Detection and Applications of Deepfake Videos', *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 8, no. 14, pp. 47–60, 2021.

[6] B. Zoph, V. Vasudevan, J. Shlens, and Q. v. Le, 'Learning Transferable Architectures for Scalable Image Recognition', 2018. doi: 10.1109/CVPR.2018.00907.

[7] Ş. Korkmaz and M. Alkan, 'Deepfake Video Detection Using Deep Learning Algorithms', *Politeknik Dergisi*, p. 1.

[8] S. Fung, X. Lu, C. Zhang, and C. T. Li, 'DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning', in *Proceedings of the International Joint Conference on Neural Networks*, 2021, vol. 2021-July. doi: 10.1109/IJCNN52387.2021.9534089.

[9] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, 'A new deep learning-based methodology for video deepfake detection using xgboost', *Sensors*, vol. 21, no. 16, 2021, doi: 10.3390/s21165413.

[10] D. Gong, Y. J. Kumar, O. S. G. Z. Ye, and W. Chi, 'DeepfakeNet, an Efficient Deepfake Detection Method', *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120622.

[11] X. Jin, D. Ye, and C. Chen, 'Countering Spoof: Towards Detecting Deepfake with Multidimensional Biological Signals', *Security and Communication Networks*, vol. 2021, 2021, doi: 10.1155/2021/6626974.

[12] H. Agarwal, A. Singh, and D. Rajeswari, 'Deepfake Detection Using SVM', 2021. doi: 10.1109/ICESC51422.2021.9532627.

[13] T. Evgeniou and M. Pontil, 'Support vector machines: Theory and applications', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2001, vol. 2049 LNAI. doi: 10.1007/3-540-44673-7_12.

[14] Y. Al-Dhabi and S. Zhang, 'Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)', 2021. doi: 10.1109/CSAIEE54046.2021.9543264.

[15] D. Patel, J. Motiani, A. Patel, and M. H. Bohara, 'DeepFake Creation and Detection Using LSTM, ResNext', in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 101, 2022. doi: 10.1007/978-981-16-7610-9_75.

[16] S. Hochreiter and J. Schmidhuber, 'Long Short Term Memory. Neural Computation', *Neural Comput*, vol. 9, no. 8, 1997.

[17] M. Taeb and H. Chi, 'Comparison of Deepfake Detection Techniques through Deep Learning', *Journal of Cybersecurity and Privacy*, vol. 2, no. 1, 2022, doi: 10.3390/jcp2010007.

[18] X. Yang, Y. Li, and S. Lyu, 'Exposing Deep Fakes Using Inconsistent Head Poses', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019, vol. 2019-May. doi: 10.1109/ICASSP.2019.8683164.

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

[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 International Conference on Computer Vision*, 2019, vol. 2019-October. doi: 10.1109/ICCV.2019.00009.

[21] B. Dolhansky *et al.*, 'The deepfake detection challenge (dfdc) dataset', *arXiv preprint arXiv:2006.07397*, 2020.

[22] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, 'Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics', 2020. doi: 10.1109/CVPR42600.2020.00327.

[23] X. Chang, J. Wu, T. Yang, and G. Feng, 'DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network', in *Chinese Control Conference, CCC*, 2020, vol. 2020-July. doi: 10.23919/CCC50068.2020.9189596.

[24] L. Deng, H. Suo, and D. Li, 'Deepfake Video Detection Based on EfficientNet-V2 Network', *Comput Intell Neurosci*, vol. 2022, 2022.