DeepFake Detection using a Deep Evolutionary Intelligence Enhanced Neural Model

Suyash Chintawar - 191IT109
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: suyash.191it109@nitk.edu.in

Naveen Shenoy - 191IT134
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: naveenshenoy.191it134@nitk.edu.in

Sarthak Jain - 191IT145
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: sarthak.191it145@nitk.edu.in

***Abstract*—With the increased popularity of social networks, digital images and videos have become very common. Facial forgery of videos aka deepfake creation has become a rising evil with more and more methods emerging to effectively morph the face of a person in a video to that of another person. Detection of such deepfakes is necessary to prevent misuse of technology. In this research, we develop neural models to detect deepfake videos with the help of genetic algorithm. Genetic algorithm is used to efficiently extract key frames from the video clips to reduce complexity and avoid stapio-temporal reduncancy within the shortlisted frames. Further processing to obtain predictions for a sequence of frames is carried out by fine-tuning InceptionV3 and EfficientNet architectures. We consider the CelebDF dataset, a widely used DeepFake benchmark, as a benchmark for this study. Experiments are performed over various variants of the EfficientNet architecture. The fine-tuned EfficientNetB4 model achieves the maximum accuracy of 96%, which is competitive with the current state-of-the-art models.**

Key words— Ant colony optimisation, feature Selection, local search, multi-label, random restructure.

I. Introduction

In recent years, the problem of deepfake, or AI-generated face-swapping videos, has gotten worse. False video is becoming more difficult to distinguish, which causes social security a lot of problems. The ongoing advancements in video quality and video manipulation technology have made deepfake detection even more challenging. Deepfake is a technique for synthesised video creation that involves swapping the face of the person in the video with the face on the provided image. As a consequence, the target person in the generated video does or says the same things as the source person. Using models like autoencoders and generative adversarial networks (GANs), it is now simple to create high-quality deep fakes thanks to advancements in deep learning approaches. While graphical image and video manipulation is sometimes used to generate deep fakes, deep learning techniques have improved to the point where it is no longer essential.These models use the source individual's facial emotions and movements from the input video to create synthetic face images of a new person exhibiting comparable expressions and movements.

Finding these modified videos is becoming more and more challenging. Security on a global scale as well as individual privacy are seriously threatened by this. Deepfakes of powerful people and leaders are produced with nefarious intentions, spreading false information and inciting violence. This could be detrimental to international relations and give the public information about the subject that they have never heard. In this work, we present a method for efficiently identifying deep fakes using deep learning and a genetic algorithm. Celeb-df, a dataset we used, contains real and deep-fake synthesised videos with comparable visual quality to those shared online.

In this study, we develop novel neural architecture enhanced by evolutionary algorithms to effectively detect deepfake video clips. A novel genetic algorithm-based frame-selection algorithm is also proposed to effectively select a subset of frames which contain human faces for processing by the neural network. Typically, k equal-spaced frames are used for selection, which may lead to data loss because the procedure described above may skip frames that contain information. Because the fitness function of the GA-based algorithm is based on the examination of frame content, duplicated and under-informative frames can be eliminated. This is specially effective in eliminating temporal redundancy between the shortlisted frames. We further fine-tune the deep neural models namely InceptionV3 and various versions of the EfficientNet architecture for the final prediction task. Since the InceptionV3 and EfficientNet models were originally designed for image classification task, they are good candidates for DeepFake detection especially to compare its performance with similar models without GA based pre-processing.

The following summarises the format of the rest of the paper. In section II, we go through the recent works that have been done in the domain of deepfake detection. After describing the proposed work and model architecture in Section III, we demonstrate the experimental setup and results in

Authors	Year	Methodology	Limitations
Agarwal et al. [1]	2019	Approach tracks of movement of facial and head features and then extracts the presence of action units and applies the model i.e SVM for prediction	Used videos of very few POIs (point of interest)
Bappy et al. [2]	2019	This paper proposes an architecture that utilizes long short-term memory (LSTM) cells, and an encoder–decoder network that extracts out manipulated regions from normal ones	Fails to work for videos with low contrast.
Chen et al. [3]	2021	Three major steps. 1) Learning of subspaces (SSL) to directly extract features. 2) Feature distillation and finally 3) Classify through Ensembling.	Techniques like GA can improve accuracy further
Zhang et al. [4]	2022	A preprocessing method based on the image pixel matrix to eliminate similar images and the residual channel attention network (RCAN) to resize the scale of images.	When the image is large enough, the cropped subspace size may only cover a relatively small region

Section IV. Finally, we conclude with the inferences obtained in Section V.

II. Literature Survey

In this section we discuss various other researches recently been done in this field of DeepFake detection. The authors in [5] have used the DFDC challenge dataset released by FaceBook. Their proposed methodology includes the use of transformers like vision transformers which are combined with EfficientNet architectures of the B0 variants for extracting the features from the video frames. Unlike many other researches, they did not use any ensemble model. The evaluation metrics used here were F1-score and the Area under the curve (AUC) scores having values 88.0% and 0.951 respectively.

In [6], the researchers have used methods like frame rate reduction and augmentation of the video frames to improve the overall performance of the deep learning model. Here, the facial regions are extracted by cropping the bounding boxes obtained from MT-CNN (Multi-task Cascaded CNN) which are further augmented by using random set of image transformation techniques. EfficientNets have been used for feature extraction of the facial images. The videos are split into frames where mean of the prediction obtained on each image are considered. The best model accuracy obtained was by using the EfficientNet-B5 variant with accuracy score of 74.4% and AUC score of 0.829.

The researchers in [3] propose a lightweight model to solve the task of DeepFake detection. It is based on a principle called as SSL (Successive subspace localization) to extract facial features from the video frames. The proposed model has been experimented on various datasets including CelebDF, UADFV, etc. The methodology includes steps like preprocessing, feature distillation, ensemble classification, etc. The performance of the proposed model was at par with that of the state-of-the-art techniques. DeFakeHop++ [7] was an improvement over the previously proposed model which has a broader coverage of the facial landmark features.

The paper [2] discusses about hybrid technique which combines CNNs and LSTMs to identify the manipulated and non-manipulated regions inside an image basically identifying whether an image has forgery or not. An encoder-decoder architecture has been exploited to perform this task. The spatial features are extracted using CNN architectures in the encoder side whereas the decoder tries to generate a binary mask separating the manipulated image region with the non-manipulated regions.

The research [8] proposes a machine learning approach to solve this task of deepfake detection. The proposed approach gives an overall accuracy of 95% on the CelebDF dataset which is slightly lesser than the state-of-the-art techniques. The proposed methodology includes steps like frame extraction, face detection, image processing, etc. The work presented in [9] proposes LRCN, a family of spatially and temporally deep models which can be incorporated into many tasks across the field of computer vision. This applies to videos which have sequential inputs such as video frames.

III. Methodology

A. Frames Selection Using Genetic Algorithm

Faces from each frames are extracted using HAAR Cascade which will be here after referred as frames in the below methodology. A genetic algorithm can be used to choose significant frames. Each frame in a video can be represented as a bit in a binary string, where a bit's value of one indicates that frame has been chosen. We anticipate that offspring will perform better than their parents in terms of the selected frames. In order to do this, we need a fitness function that lets us evaluate the parents.

We compute the histogram for each frame and examine how each frame's histogram differs from those of the other frames in order to determine the fitness function for this method. Add the difference to the score if it is less than a predetermined threshold.

The likelihood that the current frame will resemble the other frames increases as score increases. Thus, its significance is reduced. The distance factor is also taken into account. To get more information, we require unique frames that are placed closer together.As a result, net fitness value is inversely correlated with distance and the logarithmic reciprocal of the above score. Using a logarithm helps when scaling a score.

We start with randomly generated strings of 25 one-pieces each because we will be using 25 frames for the framework's subsequent steps. Finding out the fitness values of all four parents is the first stage in selecting the parents for this

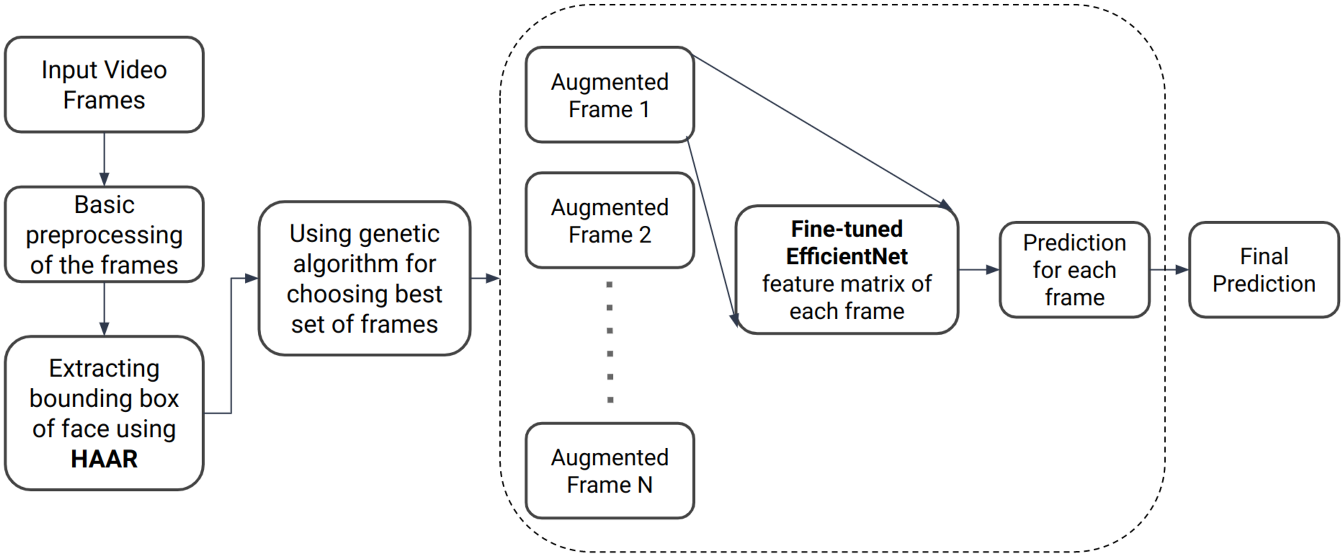
Input Video Frames
Basic preprocessing of the frames
Extracting bounding box of face using HAAR
Using genetic algorithm for choosing best set of frames
Augmented Frame 1
Augmented Frame 2
Augmented Frame N
Fine-tuned EfficientNet feature matrix of each frame
Prediction for each frame
Final Prediction

Fig. 1. Methodology

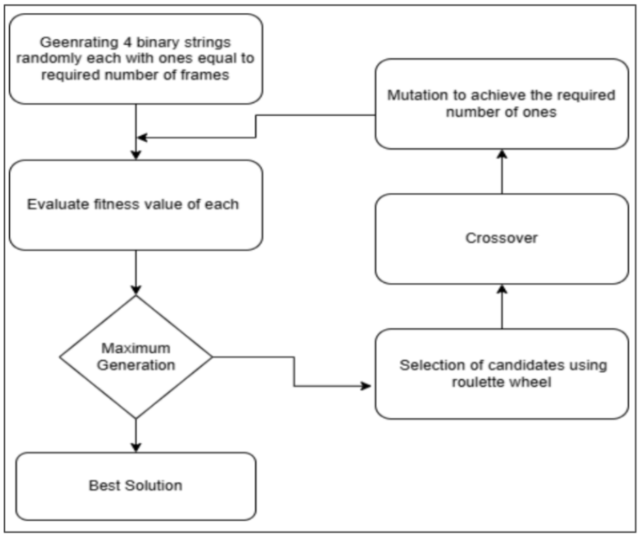
Geenrating 4 binary strings randomly each with ones equal to required number of frames
Mutation to achieve the required number of ones
Evaluate fitness value of each
Crossover
Maximum Generation
Selection of candidates using roulette wheel
Best Solution

Fig. 2. Genetic Algorithm Process

experiment. The crossover and mutation operators are then applied to the two parents that were selected using a roulette wheel.

Crossover: To perform a crossover, a random location is picked, and interchanging takes place there. Because it's possible for the final string to include more or less ones than necessary, we must perform the mutation to add a number of ones that is equal to the required number.

Mutation: If there are exactly the right number of ones, no mutation is carried out. The merging process is used if there are more ones than necessary. Up until the necessary number of ones is obtained, neighbouring segments are merged. If there are less ones than needed, a random number between 0 and 1 is chosen and transformed to a one.

Since we need 4 strings in every generation, the entire process is repeated after the mutation operations on both strings which are saved for the following generation. This keeps on till the predetermined amount of iterations. A random candidate from the current generation is chosen after fixed iterations, following which frames are chosen and are sent for the following phases. General flow is shown in Fig. 2.

B. Main Flow

Initially, from all input videos, one out of every 15 frame is extracted. Each frame is subject to basic preprocessing techniques. To apply genetic algorithm, it is necessary to extract faces from the frames. This is so that changes in the background do not distract and divert the genetic algorithm. The face bounding boxes are extracted using HAAR. Then, genetic algorithm is applied to select a subset of 15 frames from each video. The algorithm pipleine also includes inter-frame analysis to directly reject weak input frames. Each frame from the genetic algorithm selected frames is then subject to the fine-tuned EfficientNet model after data augmentation. Data augmentation includes resizing, centre cropping, random rotation, random affine transform, random horizontal and random vertical flips and finally followed by normalization. Output from the fine-tuned EfficientNet model provides the prediction for each frame of a clip. The final prediction of whether a video clip is a deepfake or not is the average of predictions over all frames of the video.

C. Fine-tuning Inception V3 model

The InceptionV3 model receives the extracted frames from the genetic algorithm as input. Initially, a pretrained InceptionV3 model is loaded. Additional layers are trained on top of the InceptionV3 model keeping the core Inception weights constant throughout the fine-tuning. Since originally, InceptionV3 is trained to predict probabilities for 1000 ImageNet classes, its output shape is (1000,). This is converted to shape of (256,) using a dense layer followed by ReLU activation. Further dropout is applied to 20% of the neurons after which a dense layer is applied to finally receive a single probability value for a frame.

D. *Fine-tuning EfficientNet models*

EfficientNet are a family of models particularly designed for the ImageNet classification task to solve the task of having huge number of training parameters in the existing models. These EfficientNet architectures proved useful for the task and surprisingly outperformed all state-of-the-art techniques making itself the best benchmark for the task. It was soon observed that EfficientNets when fine-tuned appropriately, it gave competing performance in many other computer vision tasks. In this research, we tend to exploit this power of EfficientNet models to solve the task of deep fake detection. The model architecture of EfficientNet includes inverted residual blocks which were originally proposed in MobileNetV2 [10] architectures. These enable skip connections between the ending and the beginning CNN blocks of the layers. There exists multiple variants of EfficientNet models from B0-B7. These eight models have increasing number of parameters as we move from B0 to B7. Thus, the computational complexity increases from B0 to B7. The layers include a bunch of convolutional layers combined with other layers like batch normalization, dropout, activation, zero padding, etc. The image input dimension required for each variant is different. The input size for variants changes or rather increases from 224x224x3 for EfficientNet-B0 to 600x600x3 for EfficientNet-B7. In this research we fine tune the EfficientNet variants from B0-B4 by using multiple neural network layers like Linear layers, BatchNorm layers, etc. The input shape for each model is adjusted accordingly.

IV. Experimental Results and Analysis

A. *Dataset*

The Celeb-DF dataset is used as the benchmark in this study. The Celeb-DF dataset contains 408 original YouTube clips with individuals of various ages, ethnic backgrounds, and genders, as well as 795 DeepFake videos generated from these real videos. It includes real and DeepFake generated videos with similar visual quality to those seen online. The train/test split of the dataset corresponds to 1103 training videos and 100 test videos.

B. *Preprocessing*

Here, we perform some task based data preprocessing for the EfficientNet architectures. Mainly, restructuring of the data is carried out to fit to the EfficientNet implementation. We make independent train and test directories for EfficientNet model unlike the hierarchy followed in the original raw dataset which have arranged frames of a videos according to their class labels in different directories. As we restructure the folders which originally indirectly had their class labels predefined, we create dataframes to store the label mappings of a test or train video to their respective class labels.

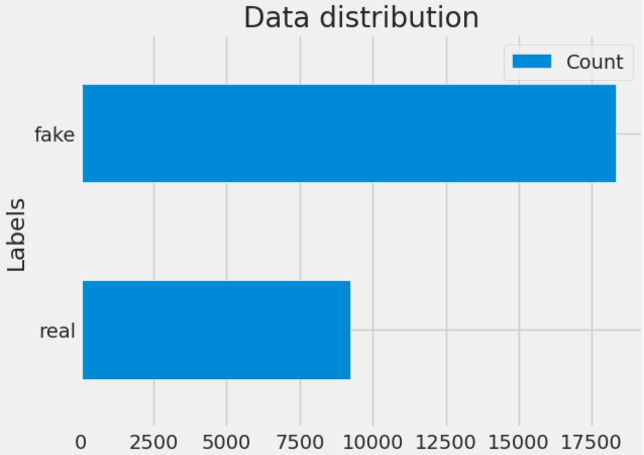

Fig. 3. Distribution of Real and DeepFake frames in the dataset

Fig. 3 shows the distribution of frames from deepfake and real videos after preprocessing. The dataset consists of more than 26k frames out of which around more than 17k are frames from deepfake video clips.

C. *Evaluation Metrics*

The evaluation metrics we use in this study is accuracy. For a given test set, accuracy is the percentage of correctly classified samples, i.e the percentage of samples rightly classified as either a deepfake or not. In tern of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), accuracy can be represented as the following:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

D. *Results and Discussion*

TABLE I
Comparison of InceptionV3 and Various EfficientNet Architectures

Sl no.	Model	Accuracy
1.	GA + Fine-tuned InceptionV3	62.0%
2.	GA + Fine-tuned EfficientNetB0	89.0%
3.	GA + Fine-tuned EfficientNetB1	93.0%
4.	GA + Fine-tuned EfficientNetB2	94.0%
5.	GA + Fine-tuned EfficientNetB3	89.0%
6.	GA + Fine-tuned EfficientNetB4	**96.0%**

Table I shows the performance of the InceptionV3 model as well as that of the various EfficientNet architectures. It is evident that all the EfficientNet architectures outperform the InceptionV3 model. Out of the EfficientNet architectures, EfficientNetB4 performs the best. There is a trend of increasing accuracy or improved performance as we move from EfficientNetB0 to EfficientNetB4. This can be attributed to the increased number of model parameters as well as the increased model complexity. The fine-tuned InceptionV3 model achieves an accuracy of 62.0%. The best EfficientNet model i.e EfficientNetB4 achieves an accuracy of 96.0%.

TABLE II
COMPARISON OF PROPOSED MODEL WITH STATE-OF-THE-ART BASELINES

Sl no.	Model	Accuracy
1.	HeadPose (Based on SVM) [11]	54.8%
2.	Meso4 (Based on CNN) [12]	53.6%
3.	Two Stream (Based on InceptionV3) [13]	55.7%
4.	DeFakeHop [3]	95.0%
5.	DeFakeHop++ [7]	97.5%
6.	GA + Fine-tuned EfficientNetB4 (proposed)	**96.0%**

Table II compares the accuracy of the best proposed model i.e GA + EfficientNetB4 with that of the current baseline and state-of-the-art models. It can be seen that both models, i.e GA + Fine-tuned InceptionV3 as well as GA + Fine-tuned EfficientNet models outperform the basic CNN based models which on based on SVM, CNN and InceptionV3 proving the superiority of our approach. Even though the GA + Fine-tuned InceptionV3 model fails to perform to the level of the recent state-of-the-art models, its improved performance over Two Stream shows the superiority of the fine-tuning as well as enhancement obtained due to genetic algorithm preprocessing. The final proposed model, i.e GA + Fine-tuned EfficientNetB4 achieves an accuracy of 96.0% which is competitive with that of the current state-of-the-art model DeFakeHop++ which obtains 97.5% accuracy.

Fig. 4 shows some sample outputs of the proposed GA + Fine-tuned EfficientNetB4 model. The frames shown are randomly selected from the residual frames after applying genetic algorithm to represent the sample. We show examples for correct classification for both deepfake and real video clips as well as incorrect classification for a real video clip.

V. CONCLUSIONS AND FUTURE WORK

In this research, we propose a novel fine-tuned neural model for deepfake detection enhanced by genetic algorithm in the preprocessing stage. Our methodology involved obtaining faces from the frames of videos using HAAR cascade and applying genetic algorithm to obtain a subset of suitable frames for further fine-tuning by InceptionV3 and EfficientNet models. Since, existing methodologies uses frames k distance apart to filter frames which leads to chances of important frames getting skipped, we design a fitness function that makes use of histograms of the frames. We find that the proposed models perform competitively when compared with the state-of-the-art baseline models. Improved performance of our approach when compared with similar deep learning models such as InceptionV3 without GA based preprocessing show superiority of our approach. Future work involves developing more task specific fine-tuned versions of the deep neural architectures.

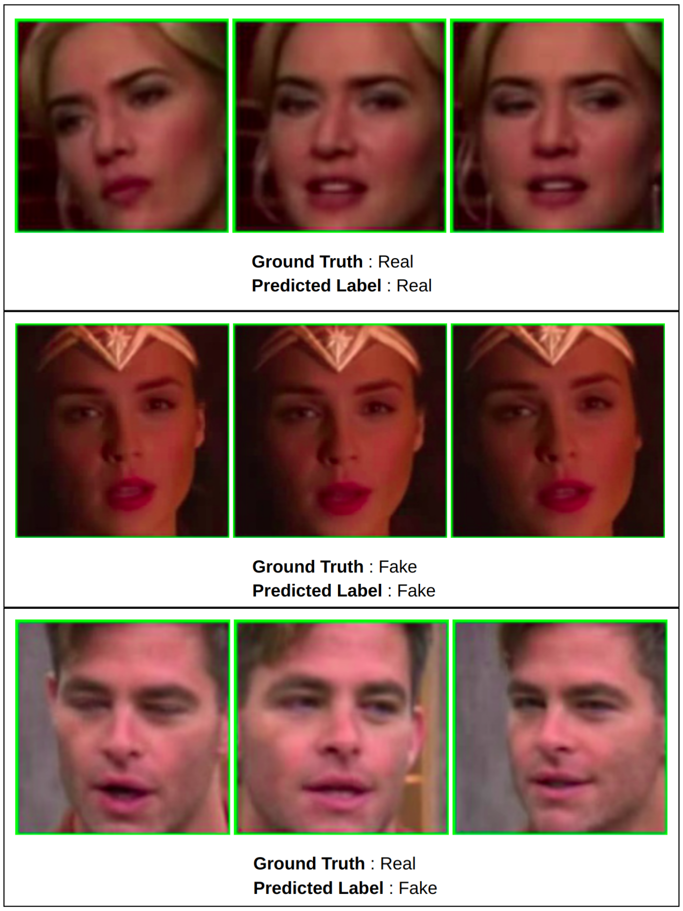
Ground Truth : Real
Predicted Label : Real
Ground Truth : Fake
Predicted Label : Fake
Ground Truth : Real
Predicted Label : Fake

Fig. 4. Samples of test inputs and their corresponding generated outputs

REFERENCES

[1] Shruti Agarwal et al. "Protecting World Leaders Against Deep Fakes." In: *CVPR workshops*. Vol. 1. 2019, p. 38.

[2] Jawadul H. Bappy et al. "Hybrid LSTM and Encoder–Decoder Architecture for Detection of Image Forgeries". In: *IEEE Transactions on Image Processing* 28.7 (2019), pp. 3286–3300. DOI: 10.1109/TIP.2019.2895466.

[3] Hong-Shuo Chen et al. "Defakehop: A light-weight high-performance deepfake detector". In: *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2021, pp. 1–6.

[4] Dengyong Zhang et al. "Cascaded-Hop For DeepFake Videos Detection". In: *KSII Transactions on Internet and Information Systems (TIIS)* 16.5 (2022), pp. 1671–1686.

[5] Davide Alessandro Coccomini et al. "Combining efficientnet and vision transformers for video deepfake detection". In: *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III*. Springer. 2022, pp. 219–229.

[6] Artem A Pokroy and Alexey D Egorov. "EfficientNets for deepfake detection: Comparison of pretrained models". In: *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. IEEE. 2021, pp. 598–600.

[7] Hong-Shuo Chen et al. "Defakehop++: An enhanced lightweight deepfake detector". In: *APSIPA Transactions on Signal and Information Processing* 11.2 (2022).

[8] Gustavo Cunha Lacerda and Raimundo Claudio da Silva Vasconcelos. "A Machine Learning Approach for DeepFake Detection". In: *arXiv preprint arXiv:2209.13792* (2022).

[9] Jeff Donahue et al. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2017), pp. 677–691. DOI: 10.1109/TPAMI.2016.2599174.

[10] Mark Sandler et al. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[11] Xin Yang, Yuezun Li, and Siwei Lyu. "Exposing deep fakes using inconsistent head poses". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 8261–8265.

[12] Darius Afchar et al. "Mesonet: a compact facial video forgery detection network". In: *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE. 2018, pp. 1–7.

[13] Peng Zhou et al. "Two-stream neural networks for tampered face detection". In: *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE. 2017, pp. 1831–1839.