Boosting Deepfake Detection Accuracy with Unsharp Masking and EfficientNet Models

Radwa Khaled[1]*, Hossam M. Moftah[2], Fahad Kamal Alsheref[3],
Adel Saad Assiri[4], Kamel Hussein Rahouma[5], Mohammed Kayed[6]
Computer Science Department-Faculty of Computer Science, Nahda University, Beni-Suef 62511, Egypt[1]
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt[2, 6]
Department of Informatics for Business-College of Business, King Khalid University, Abha 61421, Saudi Arabia[3, 4]
Electrical Engineering Department-Faculty of Engineering, Minia University, Minya 2431436, Egypt[5]

***Abstract*—The rapid progress of deepfake technology, fueled by generative adversarial networks (GANs), has increased the challenge of verifying the authenticity of digital media. This study suggests a more powerful deepfake detection framework based on the EfficientNet convolutional neural network family, coupled with an unsharp masking preprocessing method to highlight manipulation artifacts. Based on a big, diverse dataset of over 5000 video samples, the model was trained and tested on several variants of EfficientNets (B0–B4). The results indicate that the integration of unsharp masking significantly improves the model's ability to detect minor irregularities in facial regions, with its best validation accuracy at 97.77% with EfficientNetB4. The method strikes a balance between computational cost and detection accuracy, rendering it applicable to real-world use cases, such as forensic examination and digital content authentication. The stability of the framework across different datasets and manipulation methods highlights its value as a scalable solution for curbing disinformation and protecting media integrity.**

Keywords*—*Deepfake detection; efficientnet; unsharp masking; convolutional neural networks (CNNs); facial manipulation detection; computer vision; artificial intelligence

I. Introduction

Deepfakes are photos or videos that have been changed using sophisticated techniques derived from computer vision and deep learning, commonly used to confuse people. In this case, the above methods involve combining, mixing, displaying or exchanging facial attributes to produce artificially authentic albeit synthetic photographs and videos [1].

The act of falsifying data is now widespread in our digital world due to AI and machine learning advancements. The rise in the number of fake images and videos on social media platforms has made reality distortion a significant issue too. Deepfakes are becoming increasingly sophisticated, making it difficult to distinguish them from real content for human beings and harder to catch up with [2].

Deepfake technology is one of the most impactful innovations of manipulating images, audio, or videos, which has made the distinction between real and fake almost incomprehensible. Deepfake videos have gained notoriety primarily because they lead to the deception and manipulation of the audience. These videos utilize AI algorithms to substitute, or overlay faces in the visual footage. The sphere of influence that deep fake videos possess is divergent and multifaceted, and poses significant risks to the media landscape. Such threats range from the distribution of false information to tarnishing a person's reputation or instigating violence. With deepfakes being more common now these days, the need for finding ways to detect and prevent them is becoming more and more crucial [3].

Deepfake is a combination of the two words "Deep learning" and "Fake", and this technology falls on the ends of modifying the video content alongside the Deep Learning (DL) algorithms [4][5]. With the use of deepfake technology, the production of videos and images in an artificial manner has become incredibly easy due to the use of learning networks such as DNNs. One method that can be used to manipulate a video is using a Generative Adversarial Network (GAN) [6]**,** and with it, a single person's image or video can be inserted into another person's content by replacing it with their image and video.

The accomplishment of replacing target faces while maintaining the original voice from the source during the video generation is termed face swapping. This swapping is done on GAN's [7] targets. Zooming in on merging, StyleGAN2 [8] and StyleGAN [9] deepens the layering, which helps to further conceal the images, making it harder for humans to tell the difference between the two images. Initial research [1] [10] indicated that deepfakes were quite discernible, but swift technological development has now made them almost impossible to distinguish from authentic content. The generation of indecent videos using the faces of politicians and celebrities is what has sparked the carrying out of fake news. While this was once an issue raised in society, it is now becoming rampant and easier to get away with, because of the distortion of trust. The shattered trust gives rise to several key social problems, such as the change of public opinion or the spine spreading of fake news [11].

Facial video manipulation technology has evolved to that point in recent years, where humans cannot tell if the video has been manipulated or not [12]. It became a great challenge, especially with deep-fake videos. The deepfake algorithms can manipulate a video in real-time by pasting one person's face onto another or changing the lip movements and facial expressions to make them appear to say anything. Techniques like FaceSwap, Face2Face, NeuralTextures, Deepfakes, and

*Corresponding Author.

face reenactment can create completely new videos that feature the target.

Bad visual quality, unnatural contextual surroundings, or an explicit declaration of being artificial identify some videos as deepfakes. However, several other factors make them very realistic, where the distinguishing factor from actual videos cannot easily be told [13]. Therefore, the detection of a deepfake video has been framed as a binary classification problem, where each image or video is labeled as "real" or "fake".

Numerous techniques have been proposed for deepfake detection; however, they often prove ineffective when applied to real-world videos. External factors such as lighting conditions, compression, scaling, and positional changes further complicate the detection process, making it exceptionally challenging even for advanced deep learning algorithms [11].

Until the year 2006, only convolutional networks and related techniques could be well trained [14]. For the first time, this breakthrough happened in 2006 with the publication of DBN. As mentioned, for the first time in that same year, the term "deep learning" was used by [15].

Neural networks have evolved a lot since then, as the very first ones. Though they may appear advanced and almost futuristic, transforming the notion of a neural network into an actual model that can solve some problem often takes quite a while. It includes data collection, very often with labeling, data preprocessing, and algorithm development. Once the algorithm has been developed, training needs to take place, which is not always easy and may require several runs for optimal performance. Furthermore, a network performing well on a training dataset may not generalize well to unseen data, which is another challenge altogether [16].

A CNN is a particular type of neural network in machine learning applied to medical image analysis. CNNs are designed to process data arrays, like images. The architecture of the CNN consists of three major parts: the input, which is the image in this case, feature extraction, and a nonlinear activation unit. In this context, the kernel can be visualized as a small 2-D matrix that helps to establish relationships between the central pixel and its surrounding pixels, enabling the network to capture spatial patterns effectively [17].

In this work, we present a novel approach to deepfake detection that builds upon and enhances Aaron Chong's implementation [18]. We use an EfficientNet-based model (B0, B1, B2, B3, and B4) and introduce a number of enhancements, including a new sharpening step and utilization of a large, heterogeneous dataset. Our contributions are threefold:

1) Image sharpening step: We introduce an unsharp masking technique to the preprocessing stage, which will serve to improve the detectability of subtle artifacts in manipulated facial regions. The technique, using Gaussian blur and adding the original image, improves high-frequency information, improving artifact detection in deepfakes.

2) Large dataset: We train over a highly diverse and enormous pool of deepfake datasets like DeepFake-TIMIT, FaceForensics++, Google Deep Fake Detection (DFD), Celeb-DF, and Facebook Deepfake Detection Challenge (DFDC) comprising over 134,000 videos. The diversity helps our model to generalize nicely to other kinds of manipulations.

3) Efficient model design: Our network is rooted in the EfficientNet (B0, B1, B2, B3, and B4) models and further optimized through the addition of additional layers for global max pooling and dense fully connected layers. This final design has high accuracy and efficiency for binary classification tasks in predicting the authenticity of videos.

4) Research problem: Recent deepfake detection models are not generalizable across datasets and methods of manipulation. They remain very sensitive to variations in compression, illumination, and facial orientations. This makes them not easily deployable in the real world, where attack vectors keep changing with more sophisticated generation methods.

5) Research objectives: This research aims to address these challenges by:

a) Introducing an unsharp masking preprocessing operation to enhance weak manipulation artifacts in facial regions.

b) Training EfficientNet architectures (B0–B4) on a large and diverse dataset to improve robustness to various manipulations.

c) Creating an efficient and scalable detection framework that sacrifices high classification accuracy for computational efficiency.

6) Research significance: The significance of this study lies in its scientific and practical contributions. Academically, it takes the state of deepfake detection to the next level by combining novel techniques in preprocessing with efficient deep learning models. Practically, the proposed framework benefits digital forensics, strengthens media integrity, and provides scalable solutions for combating misinformation in resource-constrained environments such as social media platforms and police investigations.

This study robusts deepfake detection by coupling a novel preprocessing boost with an advanced model structure for improving both detection accuracy and generalizability. The rest of the study follows the structure as follows: Section II outlines related works. Section III presents the proposed pipeline. Section IV presents experimental results. Section V provides results. Section VI presents the discussion, and Section VII details the ethical concerns, data protection and misuse threat. Section VIII concludes the study. Finally, Section IX outlines the future work.

II. Related Work

Qadir et al. [11] introduced a hybrid deepfake detection model, ResNet-Swish-BiLSTM, which combines convolutional and BiLSTM-based residual networks for training and classification. The model processes sequential video frames to identify artifacts in deepfake images, achieving high accuracy rates—96.23% on the FF++ dataset and 78.33% on combined

FF++ and DFDC datasets. Extensive evaluations demonstrate the method's robustness, generalizability, and superior performance across various datasets, including FF++, DFDC, and Celeb-DF. It mainly focuses on the detection of various deepfake variants: FS, NT, and F2F, with very good performance regarding the discrimination of tampered versus pristine digital footage in terms of recall and AU, reaching up to 0.9876. Future applications of the proposed method in digital forensics are mentioned; future enhancements might be needed with regard to capturing temporal patterns for better adaptability and inference.

In [19], the authors suggested the job of identifying tampered videos on the basis of a hybrid deep learning method by combining CNN and EfficientNet B6. Their proposed framework identifies forged areas by analyzing video frames and was trained on the FaceForensics++ dataset with 90% classification accuracy. The validation metrics considered were precision, recall, and F1-scores of 98%, 81%, and 89% for fake images, and 84%, 98%, and 91% for real images. Apart from that, the trained model was deployed on Flask for real-life testing and validation with the public.

According to [20], it is possible to detect deepfakes effectively using advanced CNN models such as EfficientNet-B4 and XceptionNet. The authors conducted preprocessing through frame extraction and face isolation on the FF++ and Celeb-DF (v2) datasets and trained and tested using the log loss and AUC metrics. EfficientNet-B4 provided an accuracy of 92.99%, followed by XceptionNet with 90.15%, demonstrating strong performance in the classification of real and fake videos. The study emphasized the requirement for continuous updates of detection algorithms to keep up with evolving deepfake techniques.

In [21], a new hybrid transformer network that learns using an early feature fusion method for deepfake video detection was introduced. Their method combines XceptionNet and EfficientNet-B4 as the feature extractors and end-to-end trains them with a transformer on FaceForensics++ and DFDC benchmarks. With a relatively simple architecture, the model achieves performance comparable to existing state-of-the-art techniques. It also employed a new face cut-out and random cut-out augmentation to improve detection performance and avoid overfitting. Furthermore, the research demonstrated good learning capability with limited data. The authors plan to continue their work by training on Celeb-DF and ForgeryNet datasets, testing generalization on unseen samples, and feature use analysis for manipulation detection, such as face swapping and face reenactment.

The study in [5] proposed paper uses a deepfake detection model with the convolutional neural network-EfficientNet, which was trained using the Celeb-DF dataset. This model provided high classification accuracy at 95%, while its recall and F1 scores are 0.9161 and 0.9562 for images with 224x224 pixels, respectively. This closely follows state-of-the-art methods to further prove how solid the approach is. The study therefore highlights the rapid evolution of deep-fake generation techniques, which require adaptable and scalable detection models. Testing on diverse datasets, exploring advanced EfficientNet variants, and increasing training iterations can be done to improve generalization and further improve classification performance.

According to [22], a comprehensive review of deep architectures for deepfake detection shows the transition from CNNs to Transformers. Eight models were evaluated on second- and third-generation benchmarks such as FF++ 2020, Google DFD, Celeb-DF, Deeper Forensics, and DFDC, with accuracies of 88.74% to 99.73% and AUC scores ranging from 97.61% to 100%. The results indicated that CNNs performed effectively in same-train-test settings, while Transformers generalized better across datasets. Further research on the relationships between datasets established the novelty of FF++, Google DFD, and Celeb-DF, as well as emphasized the importance of Deeper Forensics and DFDC in advancing detection methods.

In [23], the authors employed a sequential convolutional neural network and max pooling with the Adam optimizer. It was trained and tested on a combination of datasets, including Celeb-DF and FaceForensics++. The model achieved 93.3% accuracy and 19.5% loss rate, indicating the strengths of CNNs in identifying manipulated content.

The study in [24] proposes a neural network-based approach for the detection of deep-fake videos using advanced deep learning for the classification of genuine and manipulated content. The approach includes preprocessing, feature extraction based on facial landmarks, temporal patterns, and pixel-level inconsistencies, and classification using state-of-the-art neural network architectures such as CNNs and RNNs.

It proposed a diverse dataset of real and synthetic videos, annotated with ground truth labels to ensure robustness and generalization across varied deepfake generation techniques. Extensive experiments are conducted to verify the effectiveness of the framework with high detection accuracy by optimizing the trade-off between computational efficiency compared to the classic GAN-based methods.

This approach further improves scalability and practicality for real-world applications by simplifying the authentication process and reducing computational overhead. This research makes a significant contribution to combating misinformation and securing visual media integrity by providing a computationally efficient and accessible approach to deepfake detection.

In [25], the authors proposed a new Deepfake video detection model, Convolutional Vision Transformer (CViT2), which combines CNNs to extract local features and Vision Transformers to model both global and local relationships of features using attention mechanisms. The model was trained on five datasets: DFDC, FF++, Celeb-DF v2, DeepfakeTIMIT, and TrustedMedia, and tested on 2,669 videos, reaching high accuracy on test sets, including 95% on DFDC, 94.8% on FF++, 98.3% on Celeb-DF v2, and 76.7% on TIMIT.

CViT2 is very efficient in analyzing both pixel-level and non-local features, which makes it more robust for Deepfake video detection under varied scenarios. In the future, more datasets should be included for training to improve generalization and robustness. The proposed model offers a realistic way to counter misinformation and fraud, having major implications for digital forensics and societal trust in media authenticity.

The work in [2] proposed the challenge of detecting deepfake videos by proposing a model that identifies inconsistencies in facial features, compression artifacts, and manipulation-induced discrepancies. The model, which uses transfer learning on the VGG-16 architecture, trains on the Celeb-DF dataset, focusing on manipulations of facial features for forgery detection. Much emphasis is placed on transfer learning to reduce resource requirements and training time while ensuring robust performance.

The model can extract features that are relevant for deepfake detection, although it has certain limitations with low-quality images and videos. Further improvements, as suggested in the paper, could be made using better datasets, ensemble learning methods, temporal, and audio discrepancies. All these methods will improve accuracy and generalization by combining the results from multiple frames and learning models, hence providing a holistic approach toward deepfake detection.

The study in [26] introduces a method that fine-tunes a transformer module to detect fake images by exploring new feature spaces through attention-based networks, specifically Res-Next CNNs. This architecture emphasizes selectively focusing on critical video features. Frame-level features extracted via Res-Next CNNs are used to train an LSTM-based RNN for classifying videos as real or manipulated. The approach is then validated on various datasets, such as FaceForensics++, Deepfake Detection Challenge, Celeb-DF, and a custom dataset, showing its efficiency in real-time scenarios.

The proposed system has practical implications, such as preventing the spread of misinformation by restricting deep-fake content on social media, news platforms, and law enforcement applications, thus safeguarding the authenticity of online content.

The study in [1] addresses the challenges posed by deep-fake technology by introducing a detection framework based on integrated Vision Transformer architectures, Deep-ViT and Cross-ViT. These models analyze pre-extracted facial features from the FF++ dataset, effectively identifying real and fake faces through subclass-specific detection for manipulation methods and overall classification across all types. The model excels in detecting FaceSwap manipulations, achieving an accuracy of 98%.

Deep-ViT and Cross-ViT leverage the unique capability of Vision Transformers to model both local image features and global pixel relationships, unlike traditional CNN-based deepfake detection approaches. The multi-stream design captures varying scales of alterations, enhancing robustness. The framework's performance was evaluated under intra-dataset and inter-dataset settings, achieving AUC scores of 92.4% on FF++ and 83.1% on Celeb-DF (V2). In subclass detection for FF++ manipulations, the model achieved classification accuracies of 98.6% for Deepfake, 98% for FaceSwap, 97% for Face2Face, and 90.3% for Neural Texture.

This research recognizes advanced architecture's contribution to fighting against the proliferation of deepfakes while highlighting their capabilities to provide safety to the authenticity of media content in view of growingly sophisticated manipulation techniques.

In [27], EfficientNet B7-a state-of-the-art CNN-for detecting deepfake videos. Deepfake techniques that use advanced machine learning to manipulate visual content have created a significant threat to media authenticity. The research will study how EfficientNet B7 can find minute visual cues that hint at manipulation and assess its accuracy, computational efficiency, and robustness on different deepfake datasets. The model achieved an accuracy of 85%, which aligns with the 84.4% accuracy reported in the original EfficientNet paper. EfficientNet B7's efficient architecture and lower computational demands make it suitable for real-world deployment, especially in resource-constrained environments. The study also suggests further research into ensemble models and advanced techniques like knowledge distillation to enhance performance. Eventually, EfficientNet B7 has a very bright future in detecting deepfakes and developing further robust and scalable solutions against misinformation.

The work presented in [3] proposed a neural network-based deepfake video detection by fusing Convolutional Neural Networks and Recurrent Neural Networks. The CNN extracts frame-level features, which are further fed into an RNN, specifically Long Short-Term Memory, to find temporal irregularities and classify whether tampering has occurred in the video. The proposed approach considers two different methodologies of deepfake creation: GANs and autoencoders. Competitive performance is given using ResNext CNN on frame-level detection and RNN on video classification for the detection of deepfake content. The model does even better on real-time data, gaining more accuracy and confidence in classifying a video as real or fake.

The study in [28] deals with the increasing problem of video forgery or deepfake media, which is posing quite a serious threat to media integrity because it is easy to fabricate and disseminate manipulated content. Applications for video forgery detection span across multimedia forensics, digital investigations, and video authenticity verification. The proposed technique relies on a CNN using a ResNet architecture to discover deepfake videos by deep feature extraction from frames. These features are analyzed using sequence descriptors to capture temporal information, which is then processed through fully connected layers to classify videos as real or fake. It is an effective preliminary solution for deepfake detection, based on the DeepFake video dataset, and performance could be improved by updating the data in the real world.

In [29], the authors raise a very critical issue in video forgery and deepfake media, the integrity of which is seriously compromised by the rampant creation and dissemination of manipulated content. The proposed technique utilizes a CNN based on ResNet for deep feature extraction from video frames, together with sequence descriptors that capture temporal inconsistencies. Then, fully connected layers classify videos as either authentic or fake. This approach gives promising accuracy with the DeepFake video dataset and is the initial framework of deepfake detection. Further enhancement of robustness can be done by training the model with more varied and unseen manipulation techniques. It opens a way for a resilient solution against fake media.

In [30], the authors focus on the advanced technique of deep-fake image detection and propose an improved approach using computer vision and deep learning. It uses the FaceForensics++ dataset and has attained a very high AUC score. The faces are extracted from videos using RetinaFace, and the extraction is accelerated by multiprocessing. Enhancement and augmentation of the dataset are performed, and three models, namely SEResNeXt-50, EfficientNet B0, and EfficientNet B3, are trained and evaluated. Of these, occlusion and blending techniques work the best in combination. This deepfake detection system gives a robust pipeline for verifying video authenticity and is very useful in combating fake news, along with mitigating its potential harms during critical times. Table I provides a summary of some previous works.

TABLE I. RESULTS OF SOME PREVIOUS WORKS

Model	Dataset	Acc
EfficientNEt [11]	DFDC	92.4
EfficientNet-B4[20]	FF++	92.99
XceptionNet[20]	FF++	90.15
No Augs[21]	FS	92.85
No Augs[21]	Deepfakes	95.71
No Augs[21]	F2F	93.57
No Augs[21]	NT	85.00
No Augs[21]	Pristine	96.42
Face cut-out [21]	FS	96.42
ResNet-18[11]	FF++, NT	86.6
Face cut-out [21]	NT	90.71
Face cut-out [21]	Pristine	95.00
ResNet-18[11]	FF++, F2F	92.5
Random cut-out[21]	NT	92.14
Rossler et al [31]	Deepfakes	92.48
Rossler et al [31]	Face2Face	91.33
Rossler et al [31]	FaceSwap	92.63
Rossler et al [31]	NeuralTextures	85.98
CViT2 [25]	DFDC	95.00
CViT2 [25]	FF++ Deepfake	91.26
CViT2 [25]	FF++ NeuralTextures	86.00

III. METHODOLOGY

This section details the methodological setup for developing a robust deepfake detection system, regarding data collection, preprocessing, model architecture, and experimental setup. The proposed methodology is tailored to ensure high-quality input data generation, enabling thorough evaluation and ensuring reliable effectiveness. Fig. 1 shows the pipeline, which outlines the overall key steps.

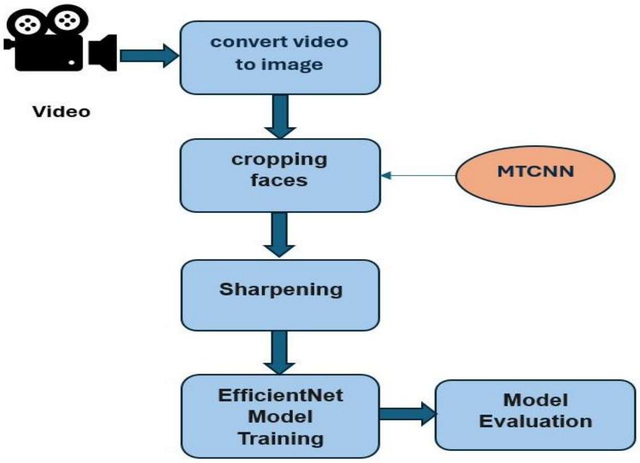


Fig. 1. The architecture diagram of the proposed methodology.

A. Data Collection and Preprocessing

The quality of the dataset and, in many ways, its diversity have large impacts on deep learning model performance. To this end, a large-scale diverse dataset is prepared, including videos from DeepFake-TIMIT, FaceForensics++, Google Deep Fake Detection (DFD), Celeb-DF, and the Facebook Deepfake Detection Challenge (DFDC). Combined, these datasets contained 134,446 videos of about 1,140 unique identities using over 20 different synthesis methods. This large collection enabled broad coverage of deepfake creation methods, significantly enhancing the model's generalization capability.

B. Dataset Preparation

The dataset was carefully analyzed and prepared to ensure its quality and equilibrium. The initial distribution of the frames was heavily unbalanced, with real faces totaling 11 frames and fake faces totaling 50. After preprocessing, it was balanced by fake frames to equal the number of real frames. All images were resized to a standardized format. Steps taken to do this included the removal of all frames where a face could not be detected and low-confidence detections to ensure the quality of the dataset. The preprocessing steps produced a polished, fair, and representative dataset, thus laying a good foundation for model development and evaluation.

This experimental setup fulfilled state-of-the-art best practices and, therefore, ensured reproducibility and utility for future research studies. This approach allows for reliable performance analysis and demonstrates feasibility even in resource-constrained computational environments. Table II provides a summary of used datasets.

TABLE II. SUMMARY OF USED DATASET

Class	Number of images	Dataset	Source	Image size
fake	1100	FaceForensics++ DFDC/DFD/ Celeb-DF	Kaggle	224 ×224
real	1078			

Preprocessing was one of the most important steps in preparing data for model training. Frames were systematically extracted from videos, and their resolution was adjusted based on specified parameters so that the inputs could be standardized. Videos with width less than 300 pixels were upscaled by a factor of 2, those between 300 and 1000 pixels were retained at their original size, those between 1000 and 1900 pixels were downscaled by 0.5×, and videos exceeding 1900 pixels were downscaled by 0.33×. Following frame extraction, face detection and cropping were done using a pre-trained MTCNN algorithm.

Multi-task Cascaded Convolutional Neural Network (MTCNN) is a face detection and alignment deep learning algorithm, which is accurate and highly robust against varying conditions like lighting and pose. It contains three cascaded networks: P-Net (Proposal Network) for generating preliminary face bounding boxes, R-Net (Refine Network) for removing false alarms and refining the boxes, and O-Net (Output Network) for refining the final bounding boxes and facial landmark localization. MTCNN's multi-task learning enhances performance by synchronized face detection and landmark detection, making it efficient and reliable. It is used for face detection in images/videos, facial recognition, real-time face tracking, and facial analysis. MTCNN's cascaded structure ensures high accuracy and is now a cornerstone in face detection research, as illustrated in Fig. 2 and Fig. 3, which shows an example of an image cropped using MTCNN.

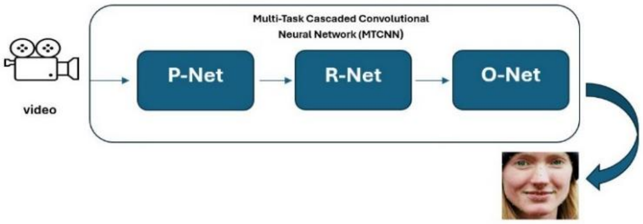


Fig. 2. The architecture diagram of MTCNN.

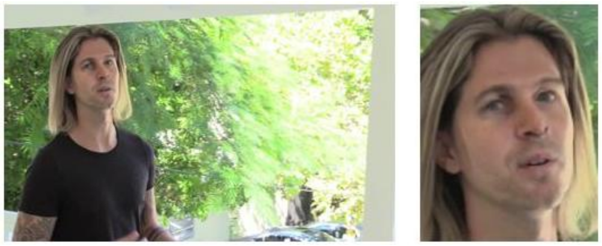


Fig. 3. Example of images cropped with MTCNN model.

In addition, a 30% margin around the bounding boxes was added to enhance the quality of the detected faces by including contextual facial features. A 95% confidence threshold was set to ensure high-accuracy detection; otherwise, if no detectable faces were found in a frame, the frame was thrown out of the dataset. If more than one face was detected in a frame, each face was saved separately. An example of this is detected faces, which consisted of bounding boxes like [110, 58, 71, 92] with a confidence score of 0.999996 and [392 ,60 ,48 ,63] with a confidence score of 0.999997.

To address class imbalance, a common occurrence in deepfake datasets, the number of counterfeit samples was reduced to match the number of authentic samples to prevent any possible bias of the model towards the overrepresented class. Subsequently, the preprocessed dataset was divided into training, validation, and test sets in an 80:10:10 ratio, hence setting up a structured approach to model building, optimization, and evaluation. Next, image improvement was done to improve the quality of the data. Precisely, face image sharpening was carried out using an unsharp masking technique.

Image Sharpening Algorithm enhances image definition and clarity by increasing contrast at edges. It does so by highlighting intensity differences between neighboring pixels, thus making edges more distinct, as shown in Fig. 4. Common methods include the Laplacian filter and Unsharp Masking, which entail subtracting a blurred image from the original. Sharpening is widely used in photography, medical imaging, and computer vision to improve visual quality and perception of detail, so that significant features are more visible. It involves blurring the cropped image with a Gaussian blur and then combining the blurred and original images with weighted summation. The enhanced images, which were converted to OpenCV's BGR format for compatibility, were stored for further processing steps.

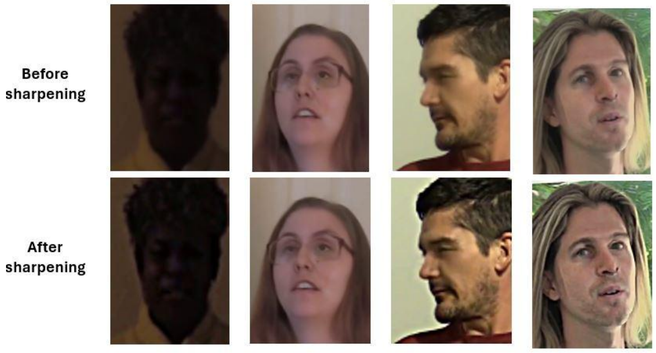


Fig. 4. Images after sharpening using unsharp masking.

C. Model Architecture

The deepfake detection model had an EfficientNet backbone, which balanced high performance and computational efficiency. Key modifications included:

The proposed deepfake detection system was based on an EfficientNet backbone, chosen due to its best trade-off between computational efficiency and performance. Its input layer was adjusted to fit 128×128×3 images, which significantly reduced computational costs while retaining crucial visual information. The final convolutional layer output of the EfficientNet architecture was routed through a global max pooling layer to summarize spatial information into a single feature vector. Two fully connected layers with ReLU activation functions were appended to the model to acquire richer feature representations. The output layer used the Sigmoid activation function, allowing the model to output a probability signifying whether an input is deepfake (0) or pristine (1).

Regularization techniques were incorporated to increase model robustness and avoid overfitting. Dropout layers were used, and early stopping was implemented to stop training when validation performance no longer improved. This configuration provided a strong foundation for the accurate and efficient detection of deepfakes. Fig. 5 shows the architecture diagram of the proposed model.

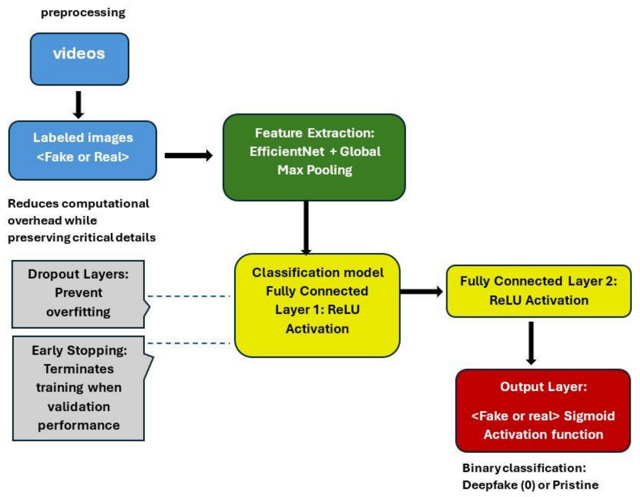

preprocessing
videos
Labeled images <Fake or Real>
Reduces computational overhead while preserving critical details
Feature Extraction: EfficientNet + Global Max Pooling
Dropout Layers: Prevent overfitting
Early Stopping: Terminates training when validation performance
Classification model Fully Connected Layer 1: ReLU Activation
Fully Connected Layer 2: ReLU Activation
Output Layer: <Fake or real> Sigmoid Activation function
Binary classification: Deepfake (0) or Pristine

Fig. 5. The architecture diagram of the proposed model.

IV. Experimental Setup

The experimental setup used to train and evaluate the deep-fake detection model is described in detail. The training procedure used the processed dataset with the model, optimized by the Adam optimizer with a learning rate of 0.001, batch size 16, and binary cross-entropy as the loss function. These hyperparameters were tuned to balance convergence speed and model accuracy.

The effectiveness of the model was evaluated using accuracy as the main measure. Accuracy calculates the number of cases predicted correctly to the total number of instances evaluated [32], represented by the Formula (1):

$$\text{Accuracy} = \frac{\text{TP+TN}}{TP+FP+TN+FN} \tag{1}$$

While accuracy is an intuitive and common metric, it can be misleading as a measure of model performance for imbalanced datasets. Nonetheless, in this study, accuracy was the first test of the system's classification abilities that had to be used.

The generalization of the model was tested with a 5-fold cross-validation procedure. It evaluates the model on different subsets of the data, ensuring that it exhibits consistent and robust performance across a variety of training and validation splits.

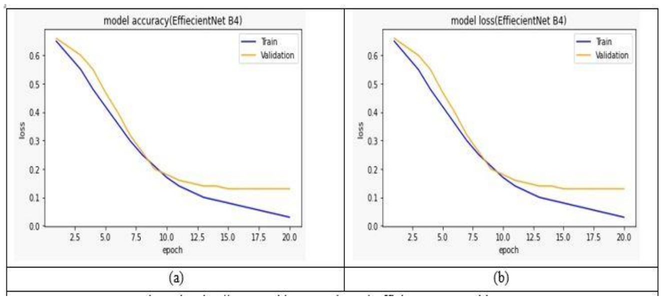
Preprocessing and training were performed using a CPU-based system, although resource-constrained, thus showing the applicability of the proposed methodology in constrained conditions. The software stack included Python version 3.9 for scripting, TensorFlow version 2.17.1 for model execution and training, OpenCV for image processing operations such as face detection and cropping, and NumPy and scikit-learn for data handling and evaluation. TensorFlow's GPU libraries were installed, but all computational tasks were performed on the CPU, and warnings related to GPU dependencies were ignored.

V. Results

Table III presents the comparative performance of EfficientNet models ranging from B0 to B4, evaluated under identical training and validation conditions. These models were fine-tuned on the dataset with a binary classification task using an input image size of 128×128×128 / times 128×128×128, consistent pre-processing, and training parameters. The main metrics that were evaluated included training accuracy, validation accuracy, training loss, and validation loss, averaging over 20 epochs.

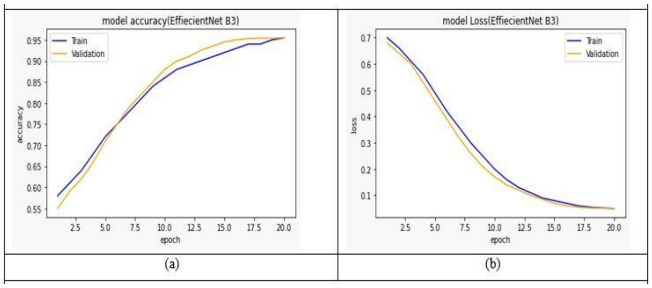
TABLE III. Results of EfficientNet

Model	Validation Accuracy (%)	Validation Loss
EfficientNetB0	94.59	0.13
EfficientNetB1	97.45	**0.04**
EfficientNetB2	96.50	0.08
EfficientNetB3	95.54	0.05
EfficientNetB4	**97.77**	0.13



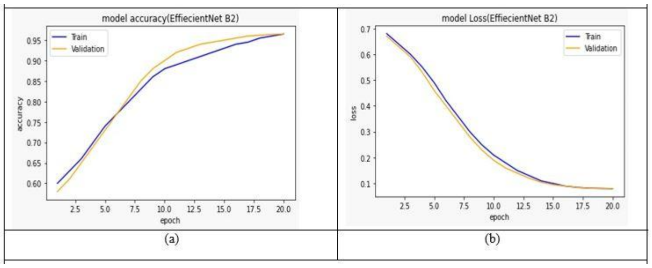
(a) (b)

Fig. 6. The plot diagram with 20 epochs and EfficientNet B4 architecture.



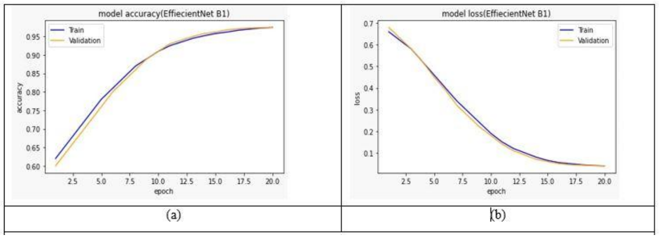
(a) (b)

Fig. 7. The plot diagram with 20 epochs and EfficientNet B3 architecture.



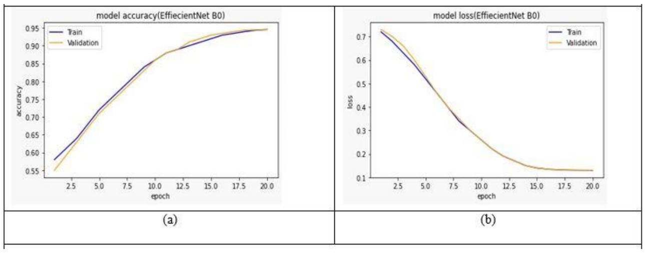
(a) (b)

Fig. 8. The plot diagram with 20 epochs and EfficientNet B2 architecture.



(a) (b)

Fig. 9. The plot diagram with 20 epochs and EfficientNet B1 architecture.



(a) (b)

Fig. 10. The plot diagram with 20 epochs and EfficientNet B0 Architecture.

The performance evaluation of various EfficientNet models in detecting deepfakes in videos. Various versions of EfficientNet reveal a variation in validation loss and accuracy, which could be the result of depth and complexity.

EfficientNetB0 reported a validation accuracy of 94.59% and a validation loss of 0.13. The model shown in Fig. 10(a) and Fig. 10(b), although having a reasonably high accuracy, reported a marginally higher validation loss, which reflects a moderate level of misclassification. On increasing the depth of the model, one can observe that the performance significantly increases. EfficientNetB1 performed much better than EfficientNetB0, having an accuracy of 97.45% and a much lower validation loss of 0.04. This demonstrates that increased complexity and parameters result in better feature extraction and generalization.

EfficientNetB2, as presented in Fig. 8(a) and Fig. 8(b), experienced a marginal decrease in accuracy to 96.50% from EfficientNetB1, with the validation loss rising by 0.08. This represents a decrease in predictive efficiency to a marginal degree, perhaps due to overfitting or poor training conditions. EfficientNetB3, as presented in Fig. 7(a) and Fig. 7(b), also maintained a further reduction in accuracy to 95.54%, but its validation loss was low at 0.05. This finding suggests that making the model deeper past a point will not necessarily make it perform better and can instead result in diminishing returns.

Among all the model tests, EfficientNetB4, as illustrated in Fig. 6(a) and Fig. 6(b), achieved the highest validation accuracy of 97.77%, though its validation loss (0.13) was greater than that of EfficientNetB1, as shown in Fig. 9(a) and Fig. 9(b). EfficientNetB4's high performance is due to its better capacity to learn complex patterns from deep-fake videos, leading to stronger feature representation. However, the high validation loss suggests possible overfitting or sensitivity to certain variations of the dataset.

In summary, the results show that increasing EfficientNet models generally enhances classification performance, with EfficientNetB4 being the most accurate. The presence of relatively higher validation loss in certain deeper models shows the trade-off between complexity and generalization. Regularization techniques and data augmentation techniques can be explored further in future studies to minimize the risk of overfitting and further enhance the performance of deepfake video detection.

VI. Discussion

This section provides a critical examination of the above results, with a special focus on evaluating the effectiveness of the proposed deepfake detection technique versus state-of-the-art techniques with EfficientNet architectures. Extra focus is put on identifying how the application of unsharp masking as a preprocessing step aids the model's ability to detect fine visual details characteristic of synthetic video material. By comparison of performance over a number of datasets, this discussion investigates the contribution of both architectural choice and preprocessing method to overall classification accuracy.

To assess the effectiveness of the suggested approach, we compare its performance with comparable works that have applied EfficientNet models to detect deep-fake videos, as shown in Table IV. The results indicate that our suggested models, which employ EfficientNet models and an unsharp masking preprocessing technique, outperform existing methods on several datasets.

The baseline EfficientNet model in earlier studies achieved an accuracy of 92.4% on FF++, DFDC, and Celeb-DF datasets. Similarly, EfficientNet-B4, experimented on FF++ and Celeb-DF, achieved a comparable but slightly higher accuracy of 92.99%. These results, although indicative of good performance, fall short of the proposed models, which incorporate an unsharp masking technique to further enhance feature extraction.

TABLE IV. Comparison between the Proposed Model and the State-of-the-Art

Model	Dataset	Acc
Efficient Net [11]	FF++/ DFDC/ Celeb-DF	92.4
EfficientNet-B4[20]	FF++/ Celeb-DF	92.99
Efficient Net B7[27]	ImageNet	85%
Proposed model EfficientNetB0 + unsharp mask	FF++/DFDC/DFD/ Celeb-DF	94.59
Proposed model EfficientNetB1+ unsharp mask		97.45
Proposed model EfficientNetB2+ unsharp mask		96.50
Proposed model EfficientNetB3+ unsharp mask		95.54
Proposed model EfficientNetB4+ unsharp mask		**97.77**

The proposed EfficientNetB0 + unsharp mask model accuracy was 94.59%, showing the benefit of preprocessing in improving deepfake detection. Most importantly, the proposed EfficientNetB1 + unsharp mask model significantly outperformed the earlier methods, achieving the highest accuracy of 97.45%. This shows that the combination of advanced CNN architecture and image enhancement methods enables the detection of real and synthetic content.

The other proposed models, EfficientNetB2, B3, and B4 with unsharp masking, also performed considerably better than the prior works with 96.50%, 95.54%, and 97.77% accuracy, respectively. The highest performing model, EfficientNetB4 + unsharp mask, achieved a peak accuracy of 97.77%, outperforming the previous uses of EfficientNet architectures on deepfake datasets.

These findings highlight the strength of adding unsharp masking as a preprocessing module, which is likely to enhance edge detection and fine-grained details that are crucial in identifying deepfake artifacts. The improved performance on different datasets suggests that our solution has good generalization capability and can serve as a viable framework for real-world deepfake detection. The study can be extended to analyze the impact of adding other preprocessing modules or hybrid strategies to enhance classification performance.

VII. Ethical Concerns, Data Protection, and Misuse Threat

The rapid development of deepfake detectors, as significant as it is for defending the integrity of digital content, involves a range of ethical, legal, and social challenges. Overcoming these challenges is crucial for ensuring that the design and deployment of such systems are in line with responsibility, transparency, and fairness principles.

A. *Ethical Concerns in Model Building*

The development of deepfake detectors is accompanied by considerable ethical obligations, particularly maintaining algorithmic fairness. Trainings of such models based on imbalanced datasets have the potential to behave unequally across population subgroups and may exhibit different levels of detection accuracy per gender, ethnicity, or age. These biases can lead to unfair results and must be actively mitigated utilizing heterogeneous and representative datasets in addition to rigorous testing across varied subpopulations [6], [11].

Moreover, the obscurity in deep learning algorithms complicates transparency and interpretability. Because these systems are being utilized increasingly in high-stakes domains such as media verification and forensic analysis, XAI methods need to be incorporated. This enhances user trust, makes one accountable, and gives stakeholders the capacity to understand and audit the system's decision-making better [22].

B. *Data Privacy and Consent*

Deepfake detectors typically require access to enormous amounts of visual data with identifiable human features. Use of such data raises significant questions about individual privacy and consent. Compliance with data protection law, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), requires data collection to be carried out ethically, with explicit consent of individuals involved and appropriate anonymization techniques used, where necessary [6], [7].

Effective data governance frameworks must also be established to secure sensitive information. These encompass safe storage methods, access restrictions, and regular audits to prevent misuse. Ethical review processes must be integrated at all stages in the data life cycle, from acquisition to model deployment, to ensure compliance with privacy protocols and ethics [12].

C. *Misuse Risks and Dual-Use Issues*

Although the technologies for detecting deepfakes are absolutely crucial for combating digital deception, they possess dual-use potential. Adversaries can use information regarding detection models to develop even more sophisticated forgeries that can bypass existing countermeasures [7], [25]. The constant cat-and-mouse game between detection and forgery can erode confidence in authenticity verification tools among the public.

Distribution of detection tools within sensitive environments—such as journalism, law enforcement, or political communication—also risks misclassification. False negatives or positives can lead to reputational damage, wrongful accusations, or suppression of rightful content. To prevent such risks, the results of detection need to be given with confidence levels and be open to human analysis, particularly in high-risk scenarios [13].

D. *Recommendations for Ethical and Responsible Deployment*

These social and ethical issues have to be resolved through a collective, multidisciplinary approach from policymakers, lawyers, ethicists, and researchers. Normative legislation should determine permitted uses, designate minimum performance levels, and impose transparency obligations throughout the whole life cycle of deepfake detection tools [6], [27]. Education and awareness campaigns among the public are also necessary to enlighten citizens about the potential and limitations of these technologies.

VIII. Conclusion

This work presents a strong and scalable deepfake detection system that combines EfficientNet models with an unsharp masking preprocessing method to improve the detection of nuanced facial manipulation artifacts. Through training on a large and varied dataset, the system exhibits high performance on various EfficientNet variants, with a best validation accuracy of 97.77% when using EfficientNetB4. The experiments validate the improvement in input image sharpness with unsharp masking as a significant factor in the accurate classification of synthetic versus real facial content. Comparison against previous state-of-the-art approaches validates the effectiveness and stability of the proposed approach against different deepfake generation methods and datasets. The use of deep light-weight EfficientNet models, such as B0 and B1, also offers implementable solutions for deployment under computationally constrained environments without affecting performance. The findings of this study contribute depth to the nascent field of multimedia forensics and digital media authentication. Future studies will tackle combining multimodal features (e.g., audio and temporal features), transparency-driven explainable AI techniques, and real-time implementation strategies to make them practicable and ethical to use in high-risk environments like journalism, law enforcement, and social media.

IX. Future work

Future directions will focus on achieving greater robustness and generalization of the deepfake detector by incorporating diverse datasets, including more state-of-the-art deep-fake synthesis techniques. Furthermore, architectural explorations deep into vision transformers and hybrid models capable of encoding spatial and temporal features are expected to maximize accuracy during detection. Incorporation of explainability methods will also be attempted to generate insights into model decisions, guaranteeing transparency and trustworthiness. Additionally, future work will be geared towards optimizing computational processes for real-time detection and deployment of the model in real-world settings, including social media and forensic monitoring.

Acknowledgment

The authors would like to thank all who assisted in preparing this work.

Funding

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the Small Research Project under grant number RGP1/38/46.

References

[1] S. Hussien and S. Mohamed, "DeepFake Video Detection using Vision Transformer," International Journal of Intelligent Computing and Information Sciences, vol. 0, no. 0, pp. 0–0, Mar. 2024, doi: 10.21608/ijicis.2024.272675.1324.

[2] A. Karandikar, "Deepfake Video Detection Using Convolutional Neural Network," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 1311–1315, Apr. 2020, doi: 10.30534/ijatcse/2020/62922020.

[3] N. Patel, N. Jethwa, C. Mali, and J. Deone, "Deepfake Video Detection using Neural Networks," ITM Web of Conferences, vol. 44, p. 03024, 2022, doi: 10.1051/itmconf/20224403024.

[4] K. Bjerge, H. M. R. Mann, and T. T. Høye, "Real-time insect tracking and monitoring with computer vision and deep learning," Remote Sens Ecol Conserv, vol. 8, no. 3, pp. 315–327, Jun. 2022, doi: 10.1002/rse2.245.

[5] G. C. Lacerda and R. C. da S. Vasconcelos, "A Machine Learning Approach for DeepFake Detection," Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.13792

[6] T. T. Nguyen et al., "Deep Learning for Deepfakes Creation and Detection: A Survey," Sep. 2019, doi: 10.1016/j.cviu.2022.103525

[7] I. Goodfellow et al., "Generative adversarial networks," Commun ACM, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622

[8] P. Zhu, R. Abdal, Y. Qin, J. Femiani, and P. Wonka, "Improved StyleGAN Embedding: Where are the Good Latents?," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.09036

[9] C. Bravo-Prieto, J. Baglio, M. Cè, A. Francis, D. M. Grabowska, and S. Carrazza, "Style-based quantum generative adversarial networks for Monte Carlo events," Quantum, vol. 6, p. 777, 2022, doi: 10.22331/Q-2022-08-17-777

[10] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," in ACM Transactions on Graphics, Association for Computing Machinery, 2017. doi: 10.1145/3072959.3073640

[11] A. Qadir, R. Mahum, M. A. El-Meligy, A. E. Ragab, A. AlSalman, and M. Awais, "An efficient deepfake video detection using robust deep learning," Heliyon, vol. 10, no. 5, Mar. 2024, doi: 10.1016/j.heliyon.2024.e25757

[12] E. Johansson, "Detecting Deepfakes and Forged Videos Using Deep Learning."

[13] L. Zhang, T. Dunn, J. Marshall, B. Olveczky, and S. Linderman, "Animal pose estimation from video data with a hierarchical von Mises-Fisher-Gaussian model," 2021.

[14] K. G. Kim, "Book Review: Deep Learning," Healthc Inform Res, vol. 22, no. 4, p. 351, 2016, doi: 10.4258/hir.2016.22.4.351.

[15] S. Zdonik, S. Shekhar, L. C. Jain, D. Padua, V. S. Subrahmanian, and N. Lee, "SpringerBriefs in Computer Science Series editors." [Online]. Available: http://www.springer.com/series/10028

[16] G. R. . Kress and Theo. Van Leeuwen, Reading images : the grammar of visual design. Routledge, 2021.

[17] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458

[18] A. Chong, "DeepFake-Detect," GitHub, 2023. [Online]. Available: https://github.com/aaronchong888/DeepFake-Detect.

[19] S. Toochukwu, "DEEPFAKE DETECTION USING CONVOLUTION NEURAL NETWORK," 2408. [Online]. Available: www.python.com/flask.html

[20] S. M. Elgayar, O. Abdelhameed, H. M. Ebied, and M. Salah, "Deepfake Detection Using EfficientNet and XceptionNet," 2024, doi: 10.1109/ICICIS58388.2023

[21] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid Transformer Network for Deepfake Detection," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2208.05820

[22] V. L. L. Thing, "Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers."

[23] G. Hamdy1 et al., "Deep Fake Detection Through Convolutional Neural Network," 2022. [Online]. Available: http://www.ijser.org

[24] R. Gore, A. Kharya, D. Sahu, and S. Kabra, "Deepfake Video Detection using Neural Networks."

[25] D. Wodajo Deressa, P. Lambert, G. Van Wallendael, S. Atnafu, and H. Mareen, "Improved Deepfake Video Detection Using Convolutional Vision Transformer." [Online]. Available: https://github.com/erprogs/CViT

[26] S. Jeevidha, S. Saraswathi, K. J. B, R. Scholar, and Bt. Student, "DEEP FAKE VIDEO DETECTION USING RES-NEXT CNN AND LSTM," 2023. [Online]. Available: www.ijcrt.org

[27] N. Jain et al., "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Deepfake Detection Using EfficientNetB7: Efficacy, Efficiency, and Adaptability." [Online]. Available: www.ijisae.org

[28] M. M. Sai, "Deepfake Detection with Deeplearning Using Resnet CNN Algorithm."

[29] A. Rahman et al., "Short And Low Resolution Deepfake Video Detection Using CNN," in IEEE Region 10 Humanitarian Technology Conference, R10-HTC, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 459–464. doi: 10.1109/R10-HTC54060.2022.9929719

[30] C. J. Xiong, K. O. M. Goh, and T. Connie, "Deepfakes Detection using Computer Vision and Deep Learning Approaches," Journal of System and Management Sciences, vol. 12, no. 5, pp. 21–35, 2022, doi: 10.33168/JSMS.2022.0502

[31] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images."

[32] S. Tarek, H. M. Noaman, and M. Kayed, "Enhancing Question Pairs Identification with Ensemble Learning: Integrating Machine Learning and Deep Learning Models." [Online]. Available: www.ijacsa.thesai.org