Ensemble of Machine Learning Classifiers for Detecting Deepfake Videos using Deep Feature

Padmashree G, and Karunakar A K

Abstract—**Rapid progress in artificial intelligence, machine learning, and deep learning over the last few decades has resulted in new methodologies and tools for altering multimedia. Deepfakes is a face-swapping technique that allows anyone to change faces in a video with incredibly realistic results. Despite its utility, if used maliciously, for example, by spreading fake news or engaging in cyberstalking, this strategy can have a substantial influence on society. This makes the identification of deepfakes a critical issue. In this paper, we propose a hybrid strategy for deepfake identification in videos that combines deep learning and machine learning. Faces are identified in the videos using YOLO-V3 face detectors and using the efficientNet deep learning model, features are extracted from the faces. Deepfakes are identified using an ensemble of machine learning classifiers such as support vector machine (SVM), decision trees(DT), k-nearest neighbor (K-NN), and naive bayes(NB) based on the max voting approach, which provides better results for datasets of varying sizes and resolutions. Experiments are carried out by integrating the Celeb-DF(v2) and FaceForensics++ (FF++) datasets and the suggested technique achieves 99.64% accuracy and proves that the suggested method is more effective than state-of-the-art methods.**

Index Terms—**Deep Learning, Deepfakes, Ensemble, Machine Learning Classifiers, EfficientNet, Face Detector, YOLO**

I. Introduction

FAKE images and videos with altered facial information, particularly those made with deepFake technologies, have recently become a big public concern[1], [2], [3], [4], [5]. The phrase "deepFake" refers to a deep learning-based approach that makes fake videos by exchanging one person's face with the face of another. Latest developments in automated video and audio editing tools, generative adversarial networks (GANs), and social media have made it possible to create and distribute high-quality tampered videos quickly[6]. Open software and mobile applications now allow anyone to create fake videos automatically, even if they have no prior expertise in the task which is even indistinguishable from human eyes.

An anonymous person on the online platform Reddit is credited with inventing deepfake technology in November 2017. The user's source code was published to GitHub in December of the same year, to allow the developer community to cooperate and further develop the idea. Deepfake technology has progressed since then, allowing for the creation of fake videos of higher and more reliable quality. Online prank efforts have become a lot of attention in recent years, due to the growing fame of social media as a means of conveying news. As social media facilitates the spread of fake news, computer vision techniques have encouraged this trend by making things simpler to create fake visuals. Malicious users are likely to utilize these fake videos to produce serious societal problems or political threats. Figures 1 and 2 show some sample videos of real and deepfake videos from Celeb-DF(v2)[7] and Face Forensics++[8] datasets respectively.

In light of this, the current study demonstrates how to integrate a deep learning technique with an ensemble of machine learning algorithms to create a highly reliable and accurate deepfake detection system. that detects visual divergences in video frames and classifies them as real or deepfakes. Using YOLO-V3 face detectors[9], faces are detected from the frames extracted from the videos, and features that help in identifying the manipulations in the videos are extracted using a fine-tuned EfficientNet architecture. Classifiers such as Support Vector Machine (SVM)[10], Decision Trees (DT)[11], K-Nearest Neighbor (KNN)[12], and Naive Bayes (NB)[13], are ensembled based on max-voting for identifying real and deepfake videos.

The main contributions of this paper include,

- Developed a hybrid approach for deepfake identification in videos that extract features using a deep-learning approach and performs classification using ensemble machine-learning classifiers.
- Faces are detected using YOLO-V3, face detectors, as faces can be detected at a faster speed with better Intersection of Union in bounding boxes, and also improve the accuracy.
- Considering the FaceForensics++ and Celeb-DF datasets, a comprehensive analysis was conducted with various deep learning and classification approaches to evaluate the effectiveness of the proposed approach.
- The usefulness of the suggested approach is proved by utilizing the CelebDF database's official cross-test assessment protocols, with promising generalization potential and new state-of-the-art results.

The rest of this paper is laid out as follows. Section II discusses a quick summary of related works. Our proposed methodology is explained in Section III. The datasets used in this investigation, as well as the performance evaluations, are described in Section IV. In Section V, we evaluate our model designs through an ablation study, and in Section VI, we visualize and analyze the results of our proposed model. Finally, in Section VII, the conclusions are presented.

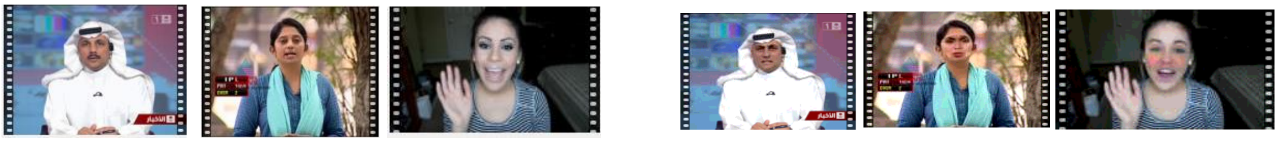
II. Related Work

Before the rise of deep learning techniques, image manipulation techniques could create realistic fake images. Nonetheless, this procedure required extremely expert image editing and, as a result, a significant amount of time. Deepfake

Manuscript received January 5, 2023; revised December 15, 2023.

Padmashree G is a research scholar in the Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India (e-mail: padmashreeg@gmail.com).

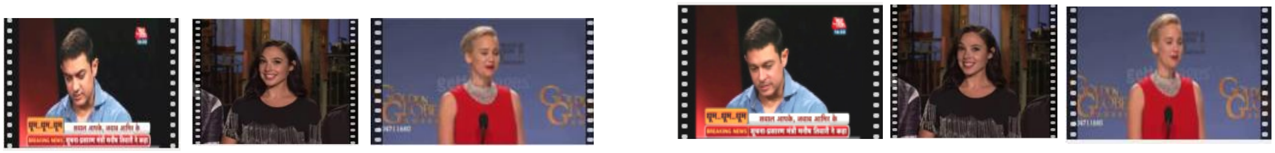
Karunakar A K is a professor in the Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India (corresponding author, e-mail:karunakar.ak@manipal.edu).



(a)

(b)

Fig. 1: FaceForensics++ Dataset (a) Real Videos (b) Fake Videos



(a)

(b)

Fig. 2: Celeb-DF(v2) Dataset (a) Real Videos (b) Fake Videos

production was made simple and quick with the help of deep learning applications such as ReFace[14], FaceApp[15], and FakeApp[16]. The DL models used in these applications, like Generative Adversarial Networks (GANs)[17] and Variational Auto-Encoders[18], have been tailored for mobile use. FSGAN[19] is one well-liked approach for producing face swapping in pictures and videos which is a GAN-based approach. FSGAN is subject-agnostic and does not require subject-specific re-training, making it an ideal tool for use on mobile devices. Furthermore, [20] provided an approach based on VAEs for disentangling and identifying feature representations in high-dimensional domains. This allows you to change someone's hair, posture, backdrop, and lighting solely based on reference materials. When creating deepfake samples, even few-shot learning is used to reduce the amount of data needed and make the process even simpler[21], [22].

Indiscernible replicas are produced by several deepfake generation models, however, these fakes can still be found using either deep learning methods or specialist scientific approaches [23], [1]. The convolution techniques used by almost all deepfake generators are visible in the image. Even though these traces can be found with a thorough investigation[23], the volume of data being published to social media sites today calls for quick and automated techniques. Deepfake images' convolutional traces can be easily found using supervised deep-learning techniques if the user supplies adequate training data. [24] presented a deep learning approach based on a relatively tiny network called MesoNet to detect whether a video is false. In their research, the authors analyze each frame individually to find deepfakes.

Additionally, [25] classifies the frames based on their optical flow. The gradient between neighboring video frames was used to train a CNN(Convolutional Neural Network) to detect sudden changes in the video. These rapid transitions demonstrate how frame-by-frame deepfake films are generated. The representation of the training dataset affects a wide range of deep learning-based deepfake detection techniques, as mentioned [26]. If a detector was not trained with data generated by a specific deepfake approach, it would almost certainly perform poorly when confronted with such a method after deployment.

The techniques for identifying and categorizing deep fakes can be categorized into two groups: Machine Learning based approaches and deep learning based frameworks. [27] developed a method to distinguish between authentic and false content in the context of traditional ML-based feature extraction techniques. Keypoint estimation in the initial phase was carried out using the Speeded up Robust Features (SURF) approach, which was then utilized to train the SVM to carry out the classification task. The blurred samples were used to assess this strategy after that. The approach [27] is effective for manipulating photos but does not generalize well to manipulating video-based multimedia content. By estimating the 3D head position from 2D facial area information, a different method for identifying changes was introduced in [28]. The estimated variance between head orientations was used as the vector of a key point in SVM training to distinguish between authentic and fake visual input. The methodology in [28] yields improved deepfake detection results; nevertheless, its effectiveness diminishes for blurred data. An approach to distinguishing the generated faces from suspicious samples was put out by [29]. With the help of the SVM and random forest, multimedia stream descriptors [30] were used to estimate key points and classify the real from the false images. The method demonstrates a low-cost method for detecting deep fakes, however for video re-encoding attacks, the detection precision decreases. The use of biological signals, such as heart rate, computed from the input video sample's facial region, as a deepfakes detection approach was first described in [7]. The computed characteristics were utilized to locate the actual and changed data using SVM and CNN-based classifiers. Although the technique is effective at detecting tampering, it is vulnerable to attacks including video post-processing. A method for detecting deep fakes was presented by [31] by computing the artificial eye-blinking pattern from the modified samples. To locate the eye-blinking patterns, the Fast-HyperFace[32] and the EAR technique (eye detection)[33] were utilized. Following the variations in eye blinks based on gender, age, behavior and time factor was then used as an integrity verification approach to discover the real and fake data samples. The technique described in [31] is effective in identifying visual manipulations, but it is less effective when applied with visual samples of people who have mental illnesses that have atypical eye-blinking movements. Because of their limited feature extraction capabilities, current machine learning-based feature extraction algorithms are incapable of dealing with post-processing threats such as the existence of extreme light fluctuations, blurring, and compression in visual data [34], [7].

The research community is assessing the effectiveness of deep learning-based algorithms for the identification of modified information to address the shortcomings of machine learning-based systems[35], [36]. A supervised learning-based method for video forensic analysis was used in [37] to demonstrate one such technique. To put it another way, the Xception network was used in conjunction with a supervised constructive loss to learn characteristics from input samples that were then categorized as original or modified. This method successfully detects deep fakes, however, the evaluation power needs to be evaluated on a more difficult dataset. In [38], another approach was proposed that used the fusion of landmarks and deep features to distinguish between actual and modified data. Though it performs better at identifying deep fakes, the work does not generalize well to dark-light visual data. To recognize the visual modifications, [39] used three different DL-based frameworks: 3D ResNet, 3D ResNeXt, and I3D. Although this approach achieves improved deep fake detection results for the 3D ResNeXt network, it struggles to generalize effectively for the testing samples that have not been seen. In [40], a different study was suggested in which deepfakes were detected using data from both frame level and temporal sequence analysis. However, this work does not perform well for the compressed video samples despite showing improved visual manipulation categorization outcomes. A technique for distinguishing between genuine and fake films was introduced by [41]. To detect the altered visual data, a two-step technique known as mask-guided identification and reconstruction was used. Deep key points were produced in the first stage and used iteratively to identify fake samples. This method suggested in is resistant to the classification of deep fakes, but it is ineffective against adversarial attacks. Furthermore, [42], proposed a technique that uses a 3D CNN approach for deepfakes detection, resulting in superior visual manipulation outcomes but with a higher computational cost. [43] presented a framework for deepfake detection and classification that utilized several pre-trained frameworks to compute key points. The SVM classifier was then trained to distinguish between genuine and fake videos using the retrieved features. The DenseNet-169 approach's the highest performance, but it has a greater computing cost.

Even though there has been a lot of work done to accurately detect deepfakes, performance still has to be improved. Existing efforts demonstrate deteriorated performance for samples subjected to adversarial attacks such as noise, compression, light fluctuations, blurring, scale and position modifications, and so on. Furthermore, while previous approaches are resistant to trained data, they perform poorly in unforeseen scenarios. Furthermore, the development of manipulated content with high realism necessitates a more precise approach to the reliable identification of fraudulent samples. In addition to these, Table I summarizes the various deepfake detection approaches in videos.

III. Proposed Methodology

In this section, we propose a methodology for classifying videos as real or fake. The architecture of our proposed method for detecting deepfake videos is illustrated in Figure 3. Initially, the input videos are pre-processed before being fed into the model. Later, the pre-processed videos are fed into a CNN model for extracting the deep features. For final prediction, many machine learning classifiers assess the derived features from the CNN model which helps in classifying the videos as real or fake.

A. Preprocessing

During the pre-processing phase, frames are extracted from videos. From these frames, faces are detected using YOLO-v3[9] face and are resized to 224 x 224. YOLO-V3 divides the image into $M \times M$ sized grid cells, with each cell attempting to locate the object in its center. The coordinate values of bounding boxes, confidence scores, and classification results for those boxes are then predicted in each grid cell. As a backbone network, the YOLO-V3 employs the darknet-53 network. Given that the scales and ratios of the anchor boxes play a key role in object detection, the anchor boxes and loss function are enhanced properly for face detection. For each scale, nine distinct anchor shapes are employed to detect faces.

B. Feature Extraction

A significant part of classification tasks is feature extraction. When features are extracted using image processing techniques, there is the possibility of misinterpretation. As a result, for deepfake classification, we employ a deep convolution neural network to extract more significant characteristics from videos. With the aid of compound coefficients, EfficientNet[58] suggests a new scaling technique that scales all depth, width, and resolution parameters consistently. Hence, the EfficientNet-B3 deep learning model was considered the base model that was trained on ImageNet weights and fine-tuned before training dismissing its final dense layer. To this architecture, two fully connected layers with 512 and 128 neurons along with the ReLU activation function were added. Each of these layers was followed by a dropout layer with a factor of 0.5 each which mainly reduces overfitting. Finally, a fully connected layer with two neurons with a SoftMax activation function is added. From the dense layer consisting of 512 neurons, $1 \times 1 \times 512$ features are provided as input to the ensemble classifier.

C. Classification Models

Features extracted from the Dense layer, which consists of 512 neurons, are supplied to an ensembled classifier after the model has been trained to evaluate the effectiveness of the suggested model. Support Vector Machine (SVM)[10], Decision Tree(DT)[11], K-Nearest Neighbors (KNN)[12], and Naive Bayes(NB)[13] are the classifiers under consideration for evaluation. In addition, an ensemble of all of these classifiers is evaluated for detecting deepfakes in videos. This section describes the several classifiers that were examined for evaluation.

TABLE I: Summary the different deepfake detection approaches in videos

Methods	Face detector	Dataset	Accuracy(%)	AUC(%)	Description
Habeeba [44]	dlib	UADFV DeepFakeDetection	92	90	Three-layer neural network+ varience of laplacian
Afchar et al.,[24]	Viola-Jones	Deepfake Face2Face	96.9, 95.3 98.4, 95.3	NA	Meso-4, MesoInception
Li & Lyu[41]	dlib	UADFV DeepfakeTIMIT	NA	98.7	ResNet50
Nguyen et al.,[45]	NA	Deepfake	97.69	NA	Capsule Network
Güera & Delp[46]	NA	HOHA	97.1	NA	CNN for frame feature extraction LSTM for temporal sequence analysis
Sabir et al.,[32]	Masks provided by Rossler et al. (2019)	FaceForensics++	96.9	NA	DenseNet + Gated Recurrent Unit(GRU)
Dang et al.,[47]	InsightFace	DFFD UADFV Celeb-DF	NA	Training: DFFD Testing: UADFV : 84.2 Testing: Celeb-DF: 64.4	XceptionNet + Attention-based layer
Li et al.,[48]	NA	Faceforensics Mesonet DeepfakeTIMIT	NA	98.3 95.5 100	Patch&Pair CNN
Kumar et al.,[49]	MTCNN	Celeb-DF	NA	99.2	XceptionNet
Khalil et al.,[50]	yolo	DFDC-P Celeb-DF	79.41 91.7	96.9 87.8	LBP+ HRNet + CapsNet
Wodajo et al.,[51]	MTCNN BlazeFace	DFDC	91.5	91	CNN + transformers
Singh et al.,[52]	MobileNet-SSD	DFDC	97.6	NA	EfficientNet-B1 + time-distributed layer + LSTM
De Lima et al.,[53]	RetinaFace	Celeb-DF	98.26	99.73	3D CNNs
Montserrat et al.,[54]	MTCNN	DFDC	91.88	NA	EfficientNet-B5 + Automatic Face Weighting layer + GRU
Ismail et al.,[55]	YOLO	CelebDF FaceForencics++	89.38	89.35	EfficientNet-B5 + Bi-LSTM
Ismail et al.,[56]	YOLO	CelebDF FaceForencics++	90.73	90.62	InceptionResNetV2 + XGBoost
Ismail et al.,[57]	YOLO	CelebDF FaceForencics++	95.56	95.53	Ameliorated XceptionNet CNN

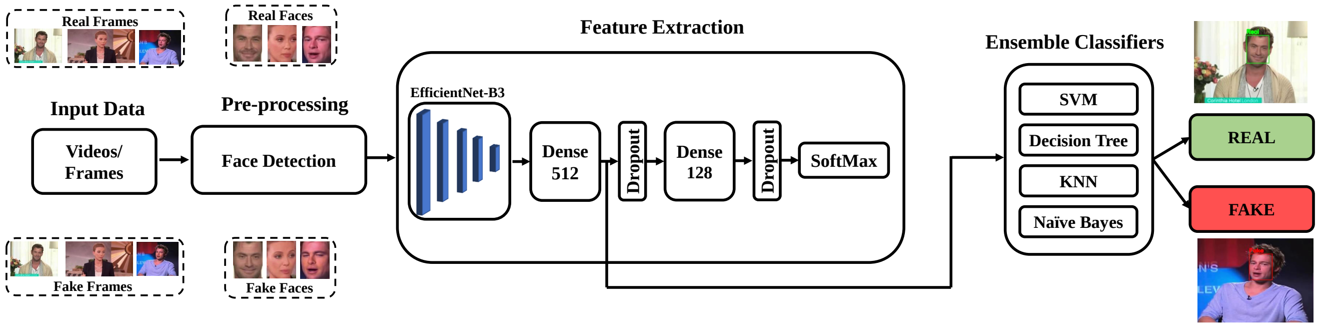


Fig. 3: Proposed model of deepfake video detection

1) *Support Vector Machine (SVM):* With a compact training area and faster processing, SVM[10] is performed utilizing an optimized hyperplane searching approach. It is a non-parametric method that can provide sensible decision limits and hence lessen misclassification. The best hyperplane is selected via SVM, which divides the data points into two groups. There are an infinite number of hyperplanes, and SVM will select the one with the biggest margin. That is how far the classifier is from the training points.

2) *Decision Trees (DT):* A machine learning method called DT[11] learns the relationships between independent variables and predictive indicators. To put it another way, the dataset is categorized based on values for predictor indicators that fall into one of the specified groups. Generally

speaking, a tree is made up of branches, leaf nodes, internal nodes, and root nodes. Each node represents an independent variable, while the branches represent alternative judgments based on the test of that predictive variable, and each leaf node is differentiated by the values of a consistent class. DT constructs a tree by arranging the dataset from the root to specific leaf nodes. The DT approach is iterative, beginning with the manipulation of the full training data set. The training dataset is divided into subsets according to the splitting rules that apply to at least one element, with the root node seeming to be the best informative element in each phase. A dividing principle can be applied by the multivariate DT to several attributes at once, but the univariate DT can only apply it to one element at a time. The DT method was run repeatedly on the training data in each branch, and the tree was finished whenever the termination requirements were met. If every training observation in a leaf node belongs to the same class, it is said to be pure.

3) K-Nearest Neighbour (KNN): Early in the 1970s, the non-parametric KNN[12] approach had its first use in mathematical applications. The main principle behind KNN is that it finds a subset of k samples in the calibration dataset that are the most similar to unknown samples (e.g., based on distance functions). The label (class) of the unknown samples can be determined by measuring the mean of the respondent parameters from these k samples, which are the class attributes in k nearest neighbors. As a result, k, the main tuning parameter of the KNN, is crucial to the KNN's effectiveness for this classifier. The parameter k was evaluated using a bootstrap methodology. To determine the ideal k value for all the training sample sets in this investigation, we examined k values between 1 and 10 and found K=5 as the best value.

4) Naïve Bayes (NB): NB[13] classifier is a probabilistic classifier based on the Bayes theorem. We find the probability of a given set of inputs for all possible values of the class variable and pick up the output with maximum likelihood.

5) Ensemble Model: The ensemble classifier model for the proposed approach is shown in Figure 4. The ensemble model using the max-voting approach is built by carefully integrating base models to create a robust model for increasing model performance[13]. Max-voting is a commonly used technique in ensemble learning, where multiple machine learning models are combined to make a prediction. In max-voting, each model in the ensemble makes a prediction for a given input, and the prediction with the highest number of votes is considered the final prediction. The main reason why max-voting is used in ensemble learning is that it can help to reduce the error rate and increase the overall accuracy of the prediction. By combining multiple models, the ensemble can capture a wider range of patterns and features in the data, leading to more robust and accurate predictions. Also, it is easy to implement and does not require a large number of computational resources, which makes it an attractive option for practical applications. We train the base classifiers SVM, DT, KNN, and NB with the 512 features extracted from the deep neural network. Then every classifier casts a vote for a particular class, and the class with the most votes wins which forms the final prediction.

IV. Experiments

This section comprises a systematic evaluation of our methodology. Our approach is first benchmarked against the state-of-the-art on two widely used deepfake datasets namely Celeb-DF(v2) and FaceForensics++ (c23). We then showcase its generalizability by conducting experiments on different datasets. Additionally, we perform a series of ablation studies to assess the influence of key components.

A. *Experimental Datasets*

FaceForensics++(c23)[8] is a forensics dataset consisting of 1000 original video sequences manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap, and Neural Textures. The data has been sourced from 977 YouTube videos, and all videos contain a trackable, mostly frontal face without occlusions, enabling automated tampering methods to generate realistic forgeries. Deepfakes were evaluated for experimentation, consisting of 1000 genuine and 1000 fake videos with a compression factor of 23.

The Celeb-DF(v2)[7] dataset contains 5,639 fake videos and 890 real videos selected from interviews with 59 celebrities in diverse ethnic and age groups.

For research purposes, FaceForensics++(c23) and Celeb-DF(v2) datasets were combined for improving the generalization in detecting deepfake videos. For training the proposed model, 712 videos were selected randomly in each category(real and fake) of videos from both datasets to form a total of 2848 videos. For testing, 340 fake videos and 178 real videos from Celeb-DF(v2) datasets were selected randomly. Table 2 provides the details of the merged dataset used in the experiment.

TABLE II: Total number of real and fake videos considered for training and testing in detecting deepfakes in videos

Datasets	**FaceForensics++(c23)**		**CelebA-DF**		**Total Videos**
	Real	**Fake**	**Real**	**Fake**	
Training	712	712	712	712	2848
Testing	-	-	178	340	518

B. *Implementation details*

All the experiments are performed using HP Elite Desk 800 G4 Workstation having 24GB NVIDIA GeForce GTX 1080 GPU, 32GB RAM, and an i7 processor with 3.7 GHz speed. The algorithm is implemented using Python 3.6 and Keras Framework. For extracting the facial region from the frames of each video, we opted for the YOLO-V3 face detector, and the extraction was performed at a rate of 2 frames per second. Augmentation techniques such as horizontal flip, zoom, and rotation were applied. The model learned the parameters for deepfake classification for 30 epochs. ”Adam” was utilized as an optimizer throughout training time with a learning rate of 0.001 to get more generic outcomes. In addition, the suggested approach uses a cross-entropy loss function to assess the CNN model's efficiency.

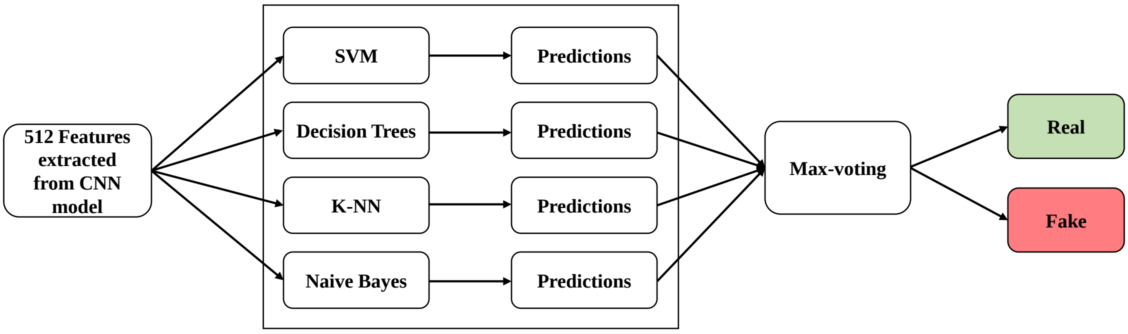


Fig. 4: Ensemble classifier model for the proposed approach

C. Model selection

In this study, we evaluated various popular deep-learning models to determine the best one for our classification task. We considered ResNet101[59], XceptionResnet-V3[60], Xception[61], and EfficientNet-B3[58] models, which are all commonly used for image classification tasks. We evaluated the models based on their accuracy, efficiency, and generalization ability. After extensive experimentation and comparison, we found that the EfficientNet-B3 model performed better than the other models in all the evaluation metrics. First, EfficientNet-B3 has demonstrated state-of-the-art performance on image classification tasks, including on the widely used ImageNet dataset. This model achieves high accuracy while being relatively computationally efficient, making it a strong candidate for our task. Second, EfficientNet-B3 has a unique architecture that effectively balances depth, width, and resolution scaling to optimize performance. This allows it to perform well on a wide range of image sizes and resolutions, making it a good fit for our diverse set of input images. Finally, we conducted experiments with all of the models under consideration and found that EfficientNet-B3 consistently outperformed the other models on our specific task, achieving higher accuracy and lower loss. Overall, the combination of high performance, computational efficiency, and adaptability to diverse image inputs made EfficientNet-B3 the best choice for our image classification task. Table III shows the testing accuracies obtained by considering the various architectures. Table IV provides the proposed model performance on the combined dataset to achieve generalization with Precision, Recall, F1-score, and Accuracy of 99.64% and AUC of 99.7%.

D. Intra test comparison

In this section, we report the results of our experiments conducted on two public datasets, FF++ and Celeb-DF, where the model is trained and tested on the same dataset to assess its ability to identify forged traces in deepfake videos. Accuracy is the chosen evaluation metric, and we provide a comprehensive visualization analysis based on our results. Our proposed method demonstrates the significant advantages of ensemble classifiers in deepfake video detection, as evident from the results presented in Table V. The use of the ”max-voting” technique further improves the performance of the model, leading to its superiority over all the compared counterparts.

E. State-of-the-Art comparison

In this study, we evaluated the effectiveness of our proposed ensemble model for detecting deepfake videos using the CelebDF-FaceForencis++ (c23) dataset. We measured the performance of the model using precision, recall, f1-score, and accuracy metrics and provided a comprehensive analysis of the results. Furthermore, we compared our model’s performance with state-of-the-art methods and observed that it outperformed other methods such as [55], [56], and [57]. Our ensemble model achieved an accuracy, precision, recall, and f1-score of 99.64%, demonstrating its superiority in detecting deepfake videos and reducing overfitting as shown in Figure 5.

V. Ablation study

We systematically evaluate our model designs through ablation studies on CelebDF-FaceForencis++ (c23) dataset. We analyze the impact of two different aspects on model performance and provide a comprehensive visualization analysis of our findings.

Study based on different face detectors In this section, we evaluate the performance of three different face detectors: dlib[67], MTCNN[68], and YOLO-V3[9], for detecting faces in videos in the context of deepfake detection. We compare the results obtained by these detectors on the CelebDF-FaceForencis++ (c23) dataset using our proposed deepfake detection model. We first preprocessed the dataset by extracting frames from each video and passing them through each face detector to obtain the detected faces. We then trained our deepfake detection model on each set of detected faces and evaluated its performance using precision, recall, F1-score, and accuracy metrics.

As shown in Figure 6, we observed that the YOLO-V3 face detector achieved the best performance with an accuracy of 99.64%, followed by dlib with an accuracy of 99.57% and MTCNN with an accuracy of 99.35%. The YOLO-V3 detector also achieved the highest precision, recall, and F1-score values among the three detectors. These results suggest that YOLO-V3 is the most effective face detector for deepfake detection among the three evaluated detectors.The superior performance of the YOLO-V3 detector can be attributed to its ability to accurately detect faces even in low-resolution and blurry frames, which are common in deepfake videos. The dlib detector also performed well, but it tends to miss some faces in certain frames. The MTCNN detector, on the other hand, struggled to detect faces in some frames,

TABLE III: Comparison of proposed deepfake detection methods with various state-of-the-art architectures

Methods	Testing Accuracy(%)
YOLO-V3 + ResNet101[59]+ Ensemble Classifier	97.48
YOLO-V3 + InceptionResnet-V3[60] + Ensemble Classifier	99.48
YOLO-V3 + Xception[61] + Ensemble Classifier	97.85
YOLO-V3 + EfficientNet-B3[58] + Ensemble Classifier (Proposed Method)	**99.64**

TABLE IV: Proposed model performance

Method	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	AUC(%)
YOLO-V3 + EfficientNet-B3[58] + Ensemble Classifier	99.64	99.64	99.64	99.64	99.7

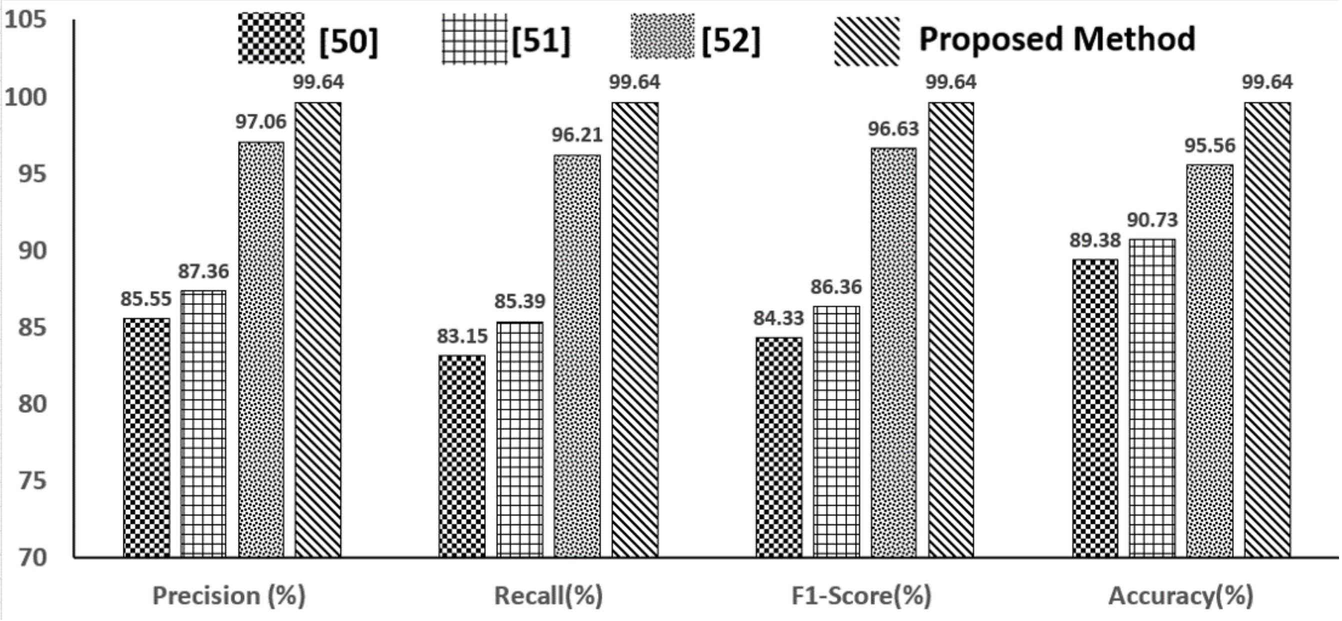


Fig. 5: Performance evaluation of deepfake detection model compared to state-of-the-art techniques

TABLE V: Intra dataset comparison of deepfake video detection

Methods	FF++	Celeb-DF
Meso4[24]	84.7	54.8
MesoInception4[24]	83	53.6
FWA[41]	80.1	56.9
DSP-FWA[41]	93	64
Multi-task[62]	76.3	54.3
Capsule[63]	96.6	57.5
Embedding[64]	99.7	66
Ensemble CNN[65]	98.67	96.89
3DCNN[66]	94.78	95.44
Proposed Method	**99.82**	**98.5**

resulting in lower overall performance. After comparing the performance of three different face detectors (dlib, MTCNN, and YOLO-V3) for deepfake detection, we conclude that YOLO-V3 outperforms the other two detectors.

Study based on individual classifiers In this section, we evaluate the performance of four individual classifiers: SVM, Decision Tree, KNN, and Naive Bayes, for the task of deepfake detection. We used the CelebDF-FaceForencis++ (c23) dataset for this experiment and evaluated the classifiers using precision, recall, F1-score, and accuracy metrics. The results are shown in Figure 7. We can observe that all four classifiers achieved high accuracy in detecting deepfakes, with the decision tree classifier having the highest accuracy of 99.55%, followed closely by SVM and with an accuracy of 99.5% and 99.17%, respectively. Naive Bayes achieved an accuracy of 96.15%, which is still a relatively high accuracy score. We also observed that Naive Bayes, which is known for its simplicity and low computational cost, achieved the lowest performance among the classifiers evaluated. This is likely due to the naive assumption that all features are independent, which may not be valid in practice. These results indicate that all four classifiers can be effective in detecting deepfakes, with SVM being the most effective among them. The high performance of these individual classifiers can be useful in situations where only one classifier can be used due to computational or other constraints. However, as observed in previous sections, an ensemble of classifiers can achieve even higher performance in detecting deepfakes, indicating the usefulness of combining multiple classifiers. These findings can be used to guide the selection of appropriate classifiers for deepfake detection in future research.

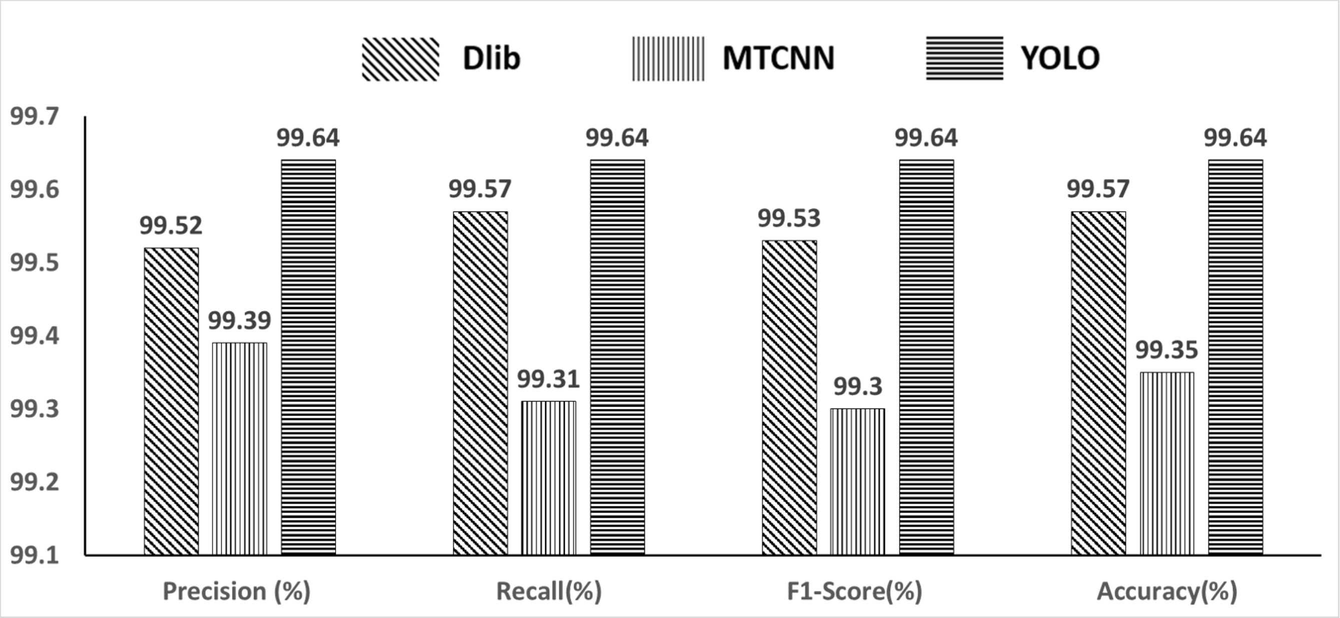


Fig. 6: Performance evaluation of deepfake detection model based on different face detectors

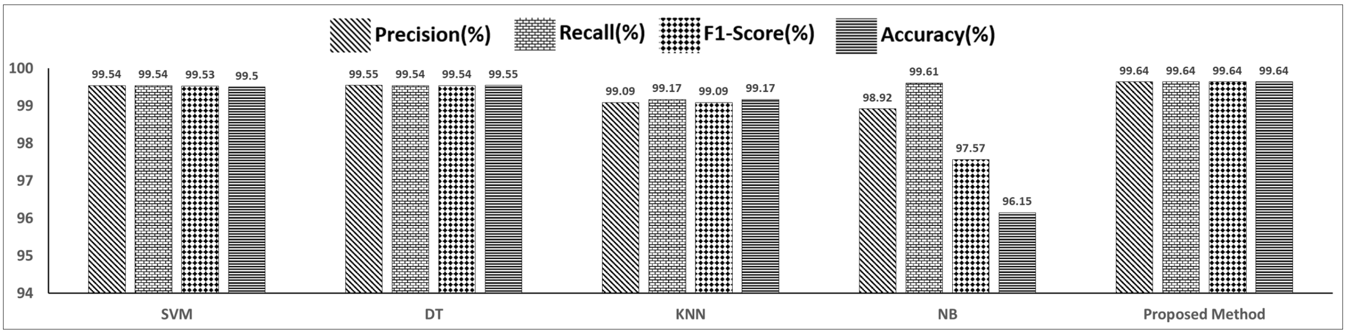


Fig. 7: Performance evaluation of deepfake detection model based on individual classifiers

VI. Visualization and analysis

In this section, we present the results of our deepfake detection model using an ensemble of machine learning classifiers. We trained the model using features extracted from the intermediate layers of a deep learning model to improve generalization. We evaluate the performance of our model using various evaluation metrics and visualization techniques. Figure 8 shows the ROC curve for our deepfake detection model. The ROC curve shows the tradeoff between the true positive rate (TPR) and the false positive rate (FPR) for different classification thresholds. The AUC score for our model is 0.997, indicating high performance in distinguishing between real and fake videos. We trained an ensemble of classifiers, including SVM, decision tree, KNN, and Naive Bayes classifiers, on the extracted features and evaluated their performance on the test set. The confusion matrix for the ensemble classifier is shown in Figure 9. The matrix indicates that the ensemble classifier achieved a high overall accuracy of 99.64%, with 99.7% precision for deepfake images and 99.4% precision for real images. We randomly selected 5 real images and 5 deepfake images from our test set and applied our ensemble classifier to predict their labels. The predicted labels for each image are shown in Figure 10. In conclusion, our deepfake detection model using features extracted from the intermediate layers of a deep learning model and an ensemble of machine learning classifiers achieved high performance in detecting deepfake videos. The ROC curve and AUC score demonstrate the model's ability to distinguish between real and fake videos, and the confusion matrix shows high accuracy in classification. The sample predicted output confirms the model's ability to correctly identify real and fake videos.

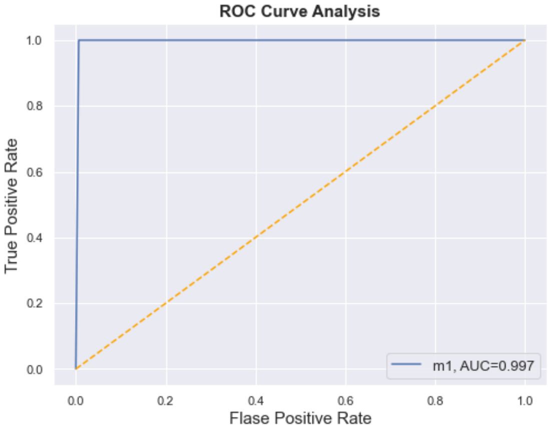


Fig. 8: AUC curve of the proposed model

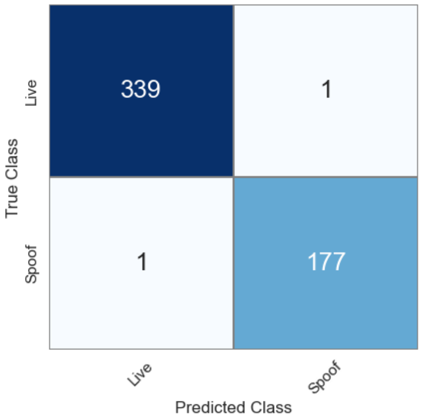


Fig. 9: Confusion matrix for our deepfake detection model

VII. Conclusion

Given the importance of videos in everyday life and online communication, being able to detect modified content in a video is extremely crucial. As a result, our research focuses on detecting face changes in video sequences using deepfake technology. While most related work on deepfakes detection highlights the performance of a single innovative technique or method, this work compares the performances of an ensemble of machine learning classifiers. The proposed ensemble technique is found to be a viable solution for the goal of face manipulation detection in videos when tested on two publicly available datasets. The suggested model demonstrated an accuracy of 99.64% utilizing the YOLO-V3 face detector, beating the deep learning-based models and providing a strong foundation for creating an efficient deepfake detector.

References

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[2] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[3] R. Cellan-Jones, "Deepfake videos double in nine months," https://www.bbc.com/news/technology-49961089, BBC, 2019.

[4] D. Citron, "How deepfake undermine truth and threaten democracy," https://www.ted.com, 2019.

[5] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools and Applications*, pp. 1–18, 2022.

[6] S. Tyagi and D. Yadav, "A detailed analysis of image and video forgery detection techniques," *The Visual Computer*, pp. 1–21, 2022.

[7] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216.

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

[9] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "Yolo-face: a real-time face detector," *The Visual Computer*, vol. 37, no. 4, pp. 805–813, 2021.

[10] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[11] J. R. Quinlan, "Combining instance-based and model-based learning," in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 236–243.

[12] N. S. Altman, "An introduction to kernel and nearest-neighbor non-parametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[13] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

[14] Reface. (2020) Reface:face swap videos. https://play.google.com/store/apps/details?id=video.reface.app.

[15] FaceApp. (2017) FaceApp-AI Face Editor. https://apps.apple.com/us/app/faceapp-perfect-face-editor/id1180884341.

[16] A. Romano, "Jordan peele's simulated obama psa is a double-edged warning against fake news," https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed, 2018.

[17] G. Ian, J. Pouget-Abadie, M. Mirza, B. Xu, and D. Warde-Farley, "Generative adversarial nets." in advances in neural information processing systems," 2014.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.

[20] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6713–6722.

[21] N. Bendre, H. T. Marín, and P. Najafirad, "Learning from few samples: A survey," *arXiv preprint arXiv:2007.15484*, 2020.

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

[23] L. Guarnera, O. Giudice, C. Nastasi, and S. Battiato, "Preliminary forensics analysis of deepfake images," in *2020 AEIT International Annual Conference (AEIT)*. IEEE, 2020, pp. 1–6.

[24] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pp. 1–7.

[25] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision workshops*, 2019, pp. 0–0.

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

[27] N. Dufour and A. Gully, "Contributing data to deepfake detection research," *Google AI Blog*, vol. 1, no. 3, 2019.

[28] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "Swapped! digital face presentation attack detection via weighted local magnitude pattern," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 659–665.

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

[30] Y. Zhang, L. Zheng, and V. L. Thing, "Automated face swapping and its detection," in *2017 IEEE 2nd international Conference on Signal and Image Processing (ICSIP)*. IEEE, 2017, pp. 15–19.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[32] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[33] H. Mo, B. Chen, and W. Luo, "Fake faces identification via convolutional neural network," in *Proceedings of the 6th ACM workshop on Information Hiding and Multimedia Security*, 2018, pp. 43–47.

[34] Y. Li, X. Yang, B. Wu, and S. Lyu, "Hiding faces in plain sight: Disrupting ai face synthesis with adversarial perturbations," *arXiv preprint arXiv:1906.09288*, 2019.

[35] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, 2019.

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

[37] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1589–1604.

[38] P. Fraga-Lamas and T. M. Fernández-Caramés, "Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality," *IT Professional*, vol. 22, no. 2, pp. 53–59, 2020.

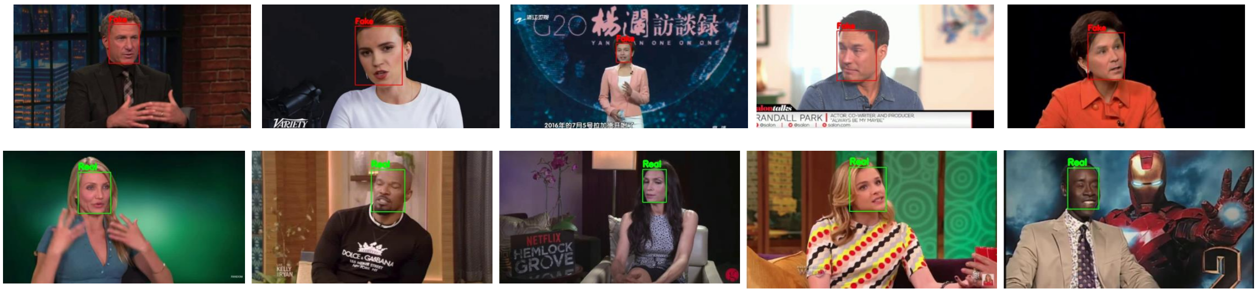


Fig. 10: Sample predicted output for our ensemble classifier. The red label indicates a deepfake image, while the green label indicates a real image.

[39] Z. Akhtar and D. Dasgupta, "A comparative evaluation of local feature descriptors for deepfakes detection," in *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, 2019, pp. 1–5.

[40] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," *arXiv preprint arXiv:1911.00686*, 2019.

[41] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[42] J. Straub, "Using subject face brightness assessment to detect 'deep fakes'(conference presentation)," in *Real-Time Image Processing and Deep Learning 2019*, vol. 10996. SPIE, 2019, p. 109960H.

[43] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7556–7566.

[44] S. Habeeba, A. Lijiya, and A. M. Chacko, "Detection of deepfakes using visual artifacts and neural network classifier," in *Innovations in Electrical and Electronic Engineering*, 2021, pp. 411–422.

[45] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2307–2311.

[46] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[47] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5781–5790.

[48] X. Li, K. Yu, S. Ji, Y. Wang, C. Wu, and H. Xue, "Fighting against deepfake: Patch&pair convolutional neural networks (ppcnn)," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 88–89.

[49] A. Kumar, A. Bhavsar, and R. Verma, "Detecting deepfakes with metric learning," in *2020 8th international Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2020, pp. 1–6.

[50] S. S. Khalil, S. M. Youssef, and S. N. Saleh, "icaps-dfake: An integrated capsule-based model for deepfake image and video detection," *Future Internet*, vol. 13, no. 4, p. 93, 2021.

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

[52] A. Singh, A. S. Saimbhi, N. Singh, and M. Mittal, "Deepfake video detection: a time-distributed approach," *SN Computer Science*, vol. 1, no. 4, pp. 1–8, 2020.

[53] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake detection using spatiotemporal convolutional networks," *arXiv preprint arXiv:2006.14749*, 2020.

[54] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu *et al.*, "Deepfakes detection with automatic face weighting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 668–669.

[55] A. Ismail, M. Elpeltagy, M. Zaki, and K. A. ElDahshan, "Deepfake video detection: Yolo-face convolution recurrent approach," *PeerJ Computer Science*, vol. 7, p. e730, 2021.

[56] 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, p. 5413, 2021.

[57] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "An integrated spatiotemporal-based methodology for deepfake detection," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21 777–21 791, 2022.

[58] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[59] S. Wu, S. Zhong, and Y. Liu, "Deep residual learning for image steganalysis," *Multimedia Tools and Applications*, vol. 77, no. 9, pp. 10 437–10 453, 2018.

[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

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

[62] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–8.

[63] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," *arXiv preprint arXiv:1910.12467*, 2019.

[64] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6111–6121, 2021.

[65] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5012–5019.

[66] Y. Wang and A. Dantcheva, "A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes," in *2020 15Th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 515–519.

[67] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[68] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

Date of modification:December 15, 2023

Brief description of the changes

Change the author name from 'Karunkar A K' to 'Karunakar A K'