Arabian Journal for Science and Engineering (2023) 48:9727–9737
https://doi.org/10.1007/s13369-022-07321-3

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Deep Fake Video Detection Using Transfer Learning Approach

Shraddha Suratkar[1] · **Faruk Kazi**[1]

Received: 19 March 2022 / Accepted: 21 September 2022 / Published online: 11 October 2022

Abstract
The usage of the internet as a fast medium for spreading fake news reinforces the requirement of computational utensils in order to fight for it. Fake videos also called deep fakes that create great intimidation in society in an assortment of social and political behaviour. It can also be utilized for malevolent intentions. Owing to the availability of deep fake generation algorithms at cheap computation power in cloud platforms, realistic fake videos or images are created. However, it is more critical to detect fake content because of the increased complexity of leveraging various approaches to smudge the tampering. Therefore, this work proposes a novel framework to detect fake videos through the utilization of transfer learning in autoencoders and a hybrid model of convolutional neural networks (CNN) and Recurrent neural networks (RNN). Unseen test input data are investigated to check the generalizability of the model. Also, the effect of residual image input on accuracy of the model is analyzed. Results are presented for both, with and without transfer learning to validate the effectiveness of transfer learning.

Keywords Deep fake detection · Convolutional neural networks (CNN) · Recurrent neural networks (RNN) · Transfer learning · Autoencoders · Residual images

1 Introduction

With the technological advancements in artificial intelligence, it has become a lot easier to create forged videos that are difficult to distinguish from reality. There is a substantial increase in fake content on the internet. Although deep fake has certain ethical applications such as in entertainment industry as well in education, malicious intents are however used to blackmail victims or to do financial frauds. Celebrities are at higher risk of being a victim of deep fakes as their images/videos are available in bulk on web. Deep fake technology is used in politics too to defame a political party right before elections to affect their votes. As the technology to generate such fake content advances, it will hardly be possible to detect these AI synthesized media in the near future. Hence, it is critical to find an approach to detect deep fakes right away.

✉ Shraddha Suratkar
sssuratkar@ce.vjti.ac.in

Faruk Kazi
fskazi@el.vjti.ac.in

[1] Department of Electrical Engineering, Veermata Jijabai Technological Institute, An Autonomous Institute, affiliated with Mumbai University, Mumbai, India

Various countermeasures using CNN have already been proposed. A considerable amount of labelled training data is the only requirement of activity recognition algorithms under extremely miscellaneous conditions. Although these measures give best results when tested on seen attacks, their accuracy drastically reduces for unseen attacks or changes in the domain. In this paper, the generalizability of convolutional autoencoder, CNN and RNN models are analyzed using transfer learning approach. These models are analysed on various benchmark datasets such as DFDC [1], Face Forensics [2], Face-Forensics++ [3], and DFD [4]. Residual image input is also explored to check its effect on generalizability. Model's accuracy is tested with and without applying transfer learning in the deep fake detection models. Various deep learning models can be explored and compared for the task of deep fake detection to select suitable CNN. The type of manipulated artifacts identified by each model can be studied to select the model for the available dataset. The contributions made by this paper are as follows:

1. Deep fake detection models are proposed using a CNN-RNN and convolutional autoencoder network.
2. The importance of LSTM in handling longer sequences of temporal data is emphasized in video-based deep fake detection.

3. Improvised the accuracy of the CNN-RNN model by rigorously training it on datasets from various distributions.
4. Analyze generalizability of CNN-RNN model and convolutional autoencoder network to detect deep fakes from a variety of datasets with unseen attacks.
5. Analyze the effect of residual image inputs on the model's accuracy.
6. Highlight the role of transfer learning and fine-tuning in improvising accuracies of video-based deep fake detection models.

The remaining part of the paper is organized as Sect. 3 explains the materials and methods of the proposed system. Section 4 provides details of the experimental setup while Sect. 5 discusses the results of the proposed system. Section 6 provides the closing remarks as well as focuses on the future research directions of this paper.

2 Related Work

Haodong Li et al. [5] demonstrated the inconsistencies in real and deep network generated in images using an efficient feature set for identifying Deep network-generated images. The presented approach provided better performance in terms of accuracy, precision and sensitivity. Here, the presence of noise affected the accuracy in detecting the fake videos.

Xinyi Ding et al. [6] made use of transfer learning to detect face-swapped images. Here, ResNet-18 was pretrained to perform object recognition on ImageNet and created a public dataset for deep fake detection work. But, the overfitting problem affected the stability of the presented approach.

Chih-Chung Hsu et al. [7] introduced deep fake detector (DeepFD) by using pairwise learning approach to improve the generalization property of the presented technique and used integrated Siamese network with the Dense Net for deep fake image detection. Thus, the presented approach detected fake videos even in the noisy environment. But, the time taken for the processing of the fake video detection was longer than conventional methods.

Mohammad Farukh Hashmi et al. [8] presented a CNN and LSTM-based deep fake detection method. Here, abnormal features obtained by comparing real and fake videos were adapted for training and these vectors were passed into Recurrent Neural Network (RNN) for evaluation of final results. The sequence of vectors was extracted from CNN and passed to the LSTM to produce the final results. But, the systems with more resources cannot be scaled up using this method.

Shruti Agarwal et al. [9] introduced phoneme-viseme mismatches method to detect deep fake videos. Three approaches were involved in the detection of fake videos manual and intensity profile. Although, the presented approaches efficiently detected the fake videos the presence of false alarms affected the novelty of the approach.

L. Minh Dang et al. [10] combined Adaptive Boosting and extreme Gradient Boosting technique and formed a hybrid framework called HF-MANFA. The presented approach had a limitation of high time consumption and memory consumption at the validation of irrelevant features.

Falko Matern et al. [11] presented a method for detecting manipulations using visual features which existed in altered videos such as the different eye color, missing reflections, etc. Here two distinct classification algorithms like logistic regression and a Multilayer Perceptron were utilized for classification. Experimental results demonstrated the superior performance of the method to the state-of-the-art approaches. But, the scheme was inefficient for high-dimensional and complex data in the visual domain.

Ekraam Sabir et al. [12] adapted CNN component RNN to extract image features along with temporal features to detect manipulations. However, better results were obtained only for public database: Face Forensics++. Thus, the dependency of the presented approach affected the performance of the approach.

Komal Chugh et al. [13] made use of Modality Dissonance Score (MDS) to compute the audio-visual dissonance and label the video as altered or original. They implemented CNN for audio stream and 3D-ResNet for visual stream.

Yuezun Li et al. [14] explored abnormal blinking pattern in forged videos to detect them as real or fake. They used the fact that there is gap of 2–10 s between each eye blink, and the length of a typical blink is between 0.1–0.4 s per blink. Long-term Recurrent Convolutional Networks (LRCN) were used for capturing the eye blinking movement.

Yuezun Li and Siwei Lyu [15] considered that the recent Deep fake generation algorithms can create images that have finite resolutions, because of which the images must be unsampled to one with the original faces in the real video. Based on this consideration they proposed a method to detect altered images.

Amirsina Torfi et al. [16] used multi-channel feature to evaluate the correlation of audio and visual signals that were mapped into a description space. Here, coupled 3D Convolutional Neural Network was utilized for the mapping procedure. But the scheme had a limitation of poor utilization of image resources.

Iain Matthews et al. [17] used HMM (hidden Markov models) methods to parameterize sequences of lip image for recognition. First, the two methods train the model using contours of outer and inner lips and used principal component analysis of shape for deriving lip reading features. After that, the third method created features from the pixel intensities using nonlinear scale-space. The experimental results demonstrated that this method performed effectively

in detecting the objects in the occluded environment. But, the poorer scalability of the system was the major drawback of this approach.

Haodong Li et al. [18] explored two approaches for detecting GAN Generated Images: intrusive and non-intrusive. In intrusive approach GAN architecture was known, the discriminator of the GAN was used to detect the fake images. In the non-intrusive approach, face quality assessment, inception scores, and latent features were investigated. But, the presented approach cannot be appropriate for detecting objects captured in the multi-camera.

Sheng-Yu Wang et al. [19] revealed that the deep fake detection models which were trained on CNN-generated images were able to generalize well on other CNN generation methods. To detect CNN-generated images, an evaluation metric and new dataset (ForenSynths) was created. However, the neural network eventually learned the background, which might produce drift and failure.

Huy H. Nguyen et al. [20] utilized multi-task learning and created a convolutional autoencoder network to concurrently detect altered videos and the altered portions for each image input. Here, an autoencoder was utilized for the detection task, followed by the Y-shaped decoder for segmenting the manipulated regions. The weights learned by classification and segmentation task enhanced the network performance by improving the generalizability of the network on matched seen attacks and unseen attacks. However, the optical flow, pose information, and deep features were not considered in this method.

Mingxing Tan et al. [21] presented a new scaling method called Compound Scaling that uniformly scaled up dimensions of CNN network, such as depth, width and input resolution. The compound's calling is based on fact that different scaling dimensions were not independent. Thus, 84.3% top-1 accuracy was obtained on ImageNet dataset for object detection. But, the major drawback of the presented approach was the classification dependency on the color of the object, ignoring its shape and texture.

The literature gap highlights failure of available systems not utilizing deep neural network to achieve required accuracy [14]. The results of the available systems utilizing deep neural networks are biased towards testing the model on same data distribution. There is need to test the models on variations of datasets to check their generalizability.

3 Materials and methods

Deep Fakes make use of the specific procedure for the effective detection of fake video approaches. Factors resembling solidity changes, and lighting differences besides the temporal discrepancies like lip and eye movements are the main factors to be considered to identify Deep Fake videos. Amongst the various techniques for Deep Fake detection, Convolution Neural Networks (CNN) is the most commonly adopted approach because of its huge capacity and scalability for applications concerning image and video processes. In CNN, features are extracted from the image followed by certain other supervised learning methods for the final classification of Deep Fake to create better and more exact models for Deep Fake Detection. However, these measures failed to provide the finest results for unseen attacks or changes in the domain. To conquer these existing challenges, the work proposed transfer learning in autoencoders and a hybrid CNN with RNN-based approach, which is depicted in Fig. 1.

Figure 1 highlights the proposed system architecture diagram. The first stage is pre-processing stage, in which frames are extracted from each input video. From each obtained frame, faces are extracted using an EfficientNet CNN model and saved so that the facial area can be worked upon instead of working on the entire frame. EfficientNet is known for its top-1% accuracy in terms of object detection as well as it is smaller than other CNN architecture. Followed by the pre-processing stage, is the training phase of deep learning models on the saved faces. The trained model can be further used as an inference engine after it reaches its best accuracy.

Upon passing a new input video to the trained model, the video passes through pre-processing stage where for each extracted face, model predicts if the frame is real or fake LSTM architecture. Finally, the average of all the values for the obtained frames is calculated to predict if the input video is a real or fake video.

3.1 CNN-based model

For this model, a combination of CNN and RNN is explored. RNN layer i.e. LSTM is added on top of CNN i.e. EfficientNet. The main important contribution of using RNN is that it takes the sequence of data into consideration in the form of feature vector as an input from the CNN. The CNN model utilized in the proposed work is EfficientNet, which is pre-trained on the ImageNet dataset. Here, the input video sequences are first passed through the convolutional layer where the features are extracted from it. Then, the features extracted are then normalized to speed up the process in the normalization layer. Then, the non-linearity in the normalized outputs is then activated in the activation layer. After that, the downsizing operation is performed in the max-pooling layer. The feature vector represents the temporal variations captured from the sequence of frames from the input video data. On every iteration, LSTM updates its cell states with the feature vectors thus distinguishing features of real and fake frames (Fig. 2). A sequence of 10 frames each with 2048 dimensional features vector is fed as an input to LSTM forming a feature vector with 10,2048 elements. Further, the size of feature vector is reduced for

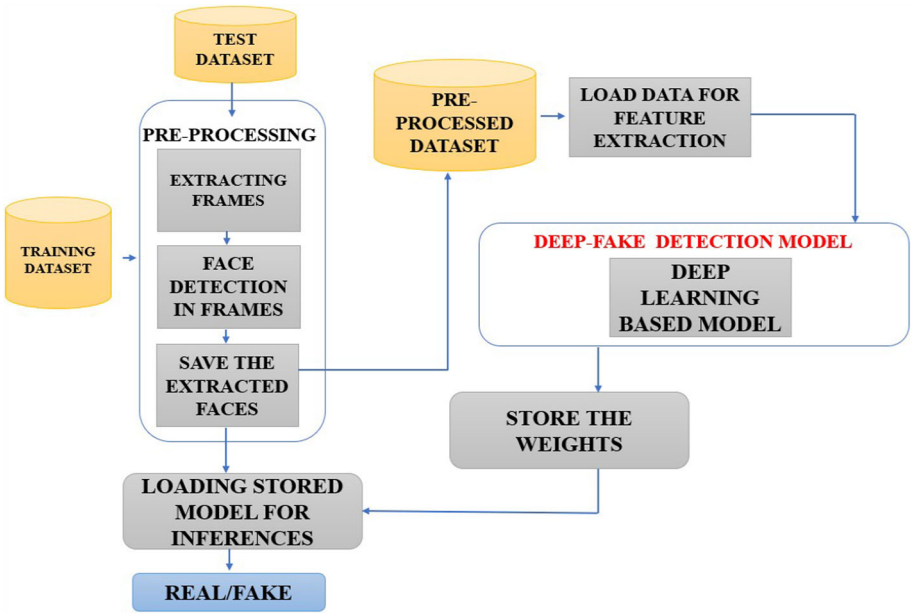

Figure1 Proposed System Architecture

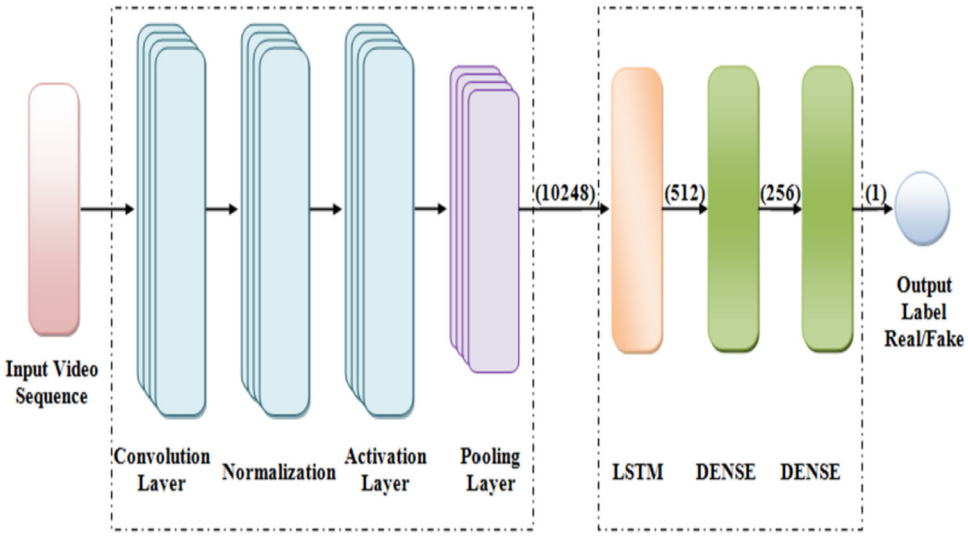

Fig. 2 Architecture of proposed CNN-based approach

efficient computations by adding dense layers. The purpose of a dense layer is to classify images based on the output of convolutional layers. LSTM being efficient in handling longer sequences of data over long span of time and memorizing the features, it is the preferred classification layer in the proposed architecture. The transfer learning approach is thus incorporated in the proposed model for the task of binary classification along with binary cross entropy loss as a metric, and the Adam optimizer approach is adapted for optimization of the model. This modification efficiently reduces the false negative values; thereby even the unseen attacks can be classified correctly and renders better classification accuracy. The architecture of the proposed CNN-based approach is shown in Fig. 2.

Let x_i be the input video sequence be detected for N number of samples. The Lossfunctionsof BinaryCrossEntropyfor the input video sequence is given by,

$$L = \frac{-1}{N} \sum_{n=1}^{N} (y_i \times \ln \left(\frac{\exp((p(y_i)))}{\sum_{i=1}^{N} ((p(y_i)z))} \right) \tag{1}$$

where y_i is the known label in the number of samples, $p(y_i)$ is the probability that point belongs to class 1 i.e., positive class, and it is calculated by:

$$z = w^T \cdot x_i \tag{2}$$

It can also be rewritten as follows

$$P(y_i = 1|x_i) = \frac{1}{1 + e^{-z}} \tag{3}$$

Further, the proposed approach utilizes an Adam optimizer to perform an optimization strategy and the steps involved in an Adam optimizer are discussed as follows.

Initially, the gradient moment (g_t) at any time (t) with respect to weight w is determined using the following expression

$$g_t = {}^{(\partial L)}/_{\partial w} w(t-1) \tag{4}$$

After that, the moving averages of the first order (m_t) and the second order moment (v_t) are revealed in Eq. (5) and (6).

$$m_t = \beta_1 m_{t-1} + (1 - \beta_2) g_t \tag{5}$$

$$v_t = \beta_2 m_{t-1} + (1 - \beta_2) g_t^2 \tag{6}$$

where β_1 and β_2 specifies the hyper moments and are initialized to zero.

Then, as very small values of moving averages are obtained by zero initialization, the bias-corrected version of the moving averages is defined as

$$\hat{m}_t = \frac{m_t}{\left(1 - \beta_1^t\right)} \tag{7}$$

$$\hat{v}_t = \frac{v_t}{\left(1 - \beta_2^t\right)} \tag{8}$$

The convergence tendency of the model is very close to the changing characteristics of the power exponential function w_t and is given as follows.

$$w_t = w_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} \tag{9}$$

For training, the model's following values are set:$\beta_1 = 0.9$, $\beta_2 = 0.99$, and the learning rate (α) is tuned differently for different models.

3.2 Proposed Method based on Autoencoder

In this approach, a Y-shaped autoencoder [20] and transfer learning are explored to analyze the generalizability of the model. The encoder is utilized for classification task and the Y shaped decoder for segmentation task. Encoder outputs latent features of the given input. These latent features are partitioned into two disjoint parts each corresponding to one of the two classes i.e. real and Fake. If the label of input image is real then corresponding part of the latent features is activated and the other half is made zero. These latent features are passed to the decoder to reconstruct and segment the image using only half of the latent features. For training, the network different loss are used viz reconstruction loss, segmentation loss and activation loss.

To measure the accuracy of dividing the latent space based on the given label, activation loss is used which is as follows,

$$L_{act} = \frac{1}{N} \sum_{i=0}^{N} |a_{i,1} - y_i| + |a_{i,0} - (1 - y_i)| \tag{10}$$

where,y_i is the known label, N is the no. of samples. $a_{i,1}$ and $a_{i,0}$ are the activation values of the corresponding halves of the latent features. L2 distance is used to calculate the reconstruction loss and it is given as follows, I_i is the original image and $I_I^{'}$ is the reconstructed image.

$$L_{rec} = \frac{1}{N} \sum_{i=0}^{N} \left\| I_i - I_i^{'} \right\| \tag{11}$$

To measure the segmentation loss, cross-entropy loss is used. Ground truth mask is g_i and segmentation out- put is s_i. It is given as:

$$L_{seg} = \frac{1}{N} \sum_{i=0}^{N} \|g_i \log(s_i) + (1 - g_i) \log(1 - s_i)\| \tag{12}$$

The total loss is calculated as the sum of all the three losses

$$L = aL_{act} + rL_{rec} + sL_{seg} \tag{13}$$

where a, r and s are the weights and their value is 1.

Input to the encoder model is an image of size (3, 256, and 256). Output of encoder network i.e. latent vector of size (128, 16, and 16) is the input to Y-shaped decoder. The output of which provides the decoded output video as real or fake.

4 Experimental Setup

4.1 Dataset Description:

In this approach, benchmark datasets such as Face- Forensics, Face Forensics++, Deep Fake Detection Challenge dataset (DFDC) and Deep Fake Detection dataset (DFD) datasets are used for training the models. Face-Forensics and Face Forensics++ dataset consists of 1000 manipulated videos, 1000 original youtube videos and 1000 binary masks for each different manipulation technique. Deep fake Detection dataset contains 3068 manipulated videos and 363 original videos from paid actors.

72% videos are used for training, 14% for validation and 14% for testing from each dataset. 200 frames are extracted from training videos whereas 20 frames were extracted from

validation and testing videos. After frame extraction faces are cropped from each frame for feature extraction.

4.2 Training

The input videos undergo pre-processing before being sent to the network. The network is first trained using manipulation technique (Re-enactment) and then tested with same as well as different manipulation techniques. Afterwards a pretrained network is used for training. The pretrained model is then fine-tuned with different manipulation techniques to check if the model is able to generalize well.

4.3 Hardware Requirement

The proposed fake video detection methodology is deployed in the working platform of NVIDIA DGX-1 with 8 V100 GPU accelerators each with 32 GB memory and runs in Linux operating system. The setup is available at CoE-CNDS laboratory of VJTI, Mumbai.

5 Results and discussion

Here, centred on disparate performance metrics, the last outcome of the proposed work with prevailing techniques was analyzed in detail. The performance analysis together with the comparative analysis is performed for proving the work's effectiveness. Here, the deep fake detection model, performed on DFDC and Face Forensics datasets is trained using various pre-trained architectures like VGG16, Inception ResNetV5, Efficient Net, and Efficient Net with LSTM and for analyzing the results, classification metrics, such as accuracy and AUC score are observed.

5.1 Performance analysis of the proposed EfficientNet

The performance of the proposed EfficientNet (CNN-based method) is evaluated based on the models trained on the different datasets, namely DFDC and Face Forensics++. Here, the performance metrics like accuracy, AUC, precision, recall, and sensitivity are utilized for evaluation purposes. The comparative analysis based on the accuracy and AUC of the proposed work is tabulated in Table 1.

Table 1 demonstrates the performance analysis of the proposed EfficientNet with various existing techniques, such as EfficientNet and LSTM, ResNetV2, and VGG16 Inception in terms of accuracy, and AUC. For this analysis, the DFDC dataset is used. For the DFDC dataset, the proposed method achieves 98.69% of accuracy and 97.26% of AUC. But the conventional method attains the accuracy and AUC at an average of 96.93% and 93.38% respectively. Generally, the higher the value of accuracy and AUC represents the better the performance of the model. Hence, when compared to the existing works, the proposed EfficientNet achieved better metrics rates. Thus, the proposed method identifies the fake videos with the highest precision rate. The clear view of Table 1 is illustrated in Fig. 3.

Table1 Performance analysis of the proposed model trained on DFDC dataset

Techniques	Metrics value (%)	
	Accuracy	AUC
Proposed EfficientNet	98.69	97.26
EfficientNet and LSTM	97.56	94.98
ResNetV2	97.64	95.18
VGG16 Inception	95.61	90

Table 2 represents the evaluation of the proposed EfficientNet with respect to accuracy and AUC. The worthiness of the model is determined by the higher rate of accuracy, and AUC. As per the statement, the proposed method achieves 85.84% of accuracy and 72.17% of AUC. But, the existing work obtains an accuracy rate that overall ranges between 69.63%-81.23%, and AUC that overall ranges between 50.48%-65.83%. This is low as compared to the proposed work. Thus, the proposed EfficientNet mitigates various complexities and enhances the robust detection of fake videos. For this analysis, the Face Forensics++ dataset is used. The clear view of Table 2 is illustrated in Fig. 4.

Figure 5 and Fig. 6 unveil the accuracy and loss of the models trained on DFDC and FF++ datasets. Epoch is nothing but the one-time processing of all images forward and backward to the network individually. The accuracy achieved on DFDC dataset is 97.7% and for Face Forensics++ dataset is 99.0%. From Fig. 5, it is clear that as the number of epochs goes on increasing the accuracy of the proposed work also increases. Meanwhile, the loss associated with the increasing epochs is decreasing and is illustrated in Fig. 6.

Figure 7 compares the training time of the proposed Efficient Net with various exist-ing techniques. From the comparative study, it is clearly known that the proposed technique takes a minimum amount of training time, such that 51329 ms is taken by the classifier to complete the training process, whereas the existing Efficient Net and LSTM, ResNetV2, and VGG16 inception take 74695 ms, 79403 ms, and 87547 ms respectively to complete the training process. Hence, the overall time of the entire model can be increased, but the proposed method completes the entire task quickly as possible, thereby the time complexity of the work can be alleviated.

Table 3 depicts the performance evaluation of the proposed real or fake video detection approach using the CNN-based

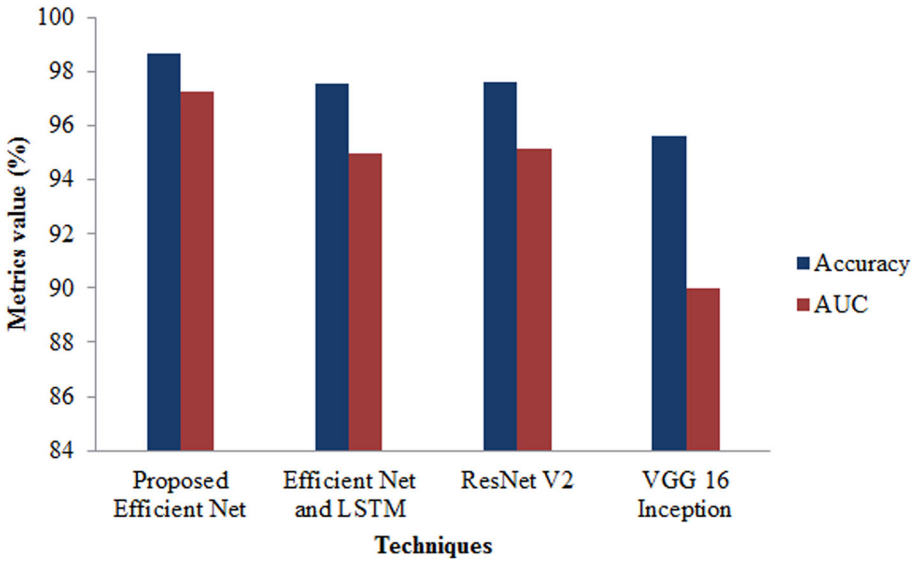

Fig. 3 Comparative analysis of the proposed model trained on DFDC dataset

and autoencoder network. From Table 3, it is transparent that, the proposed approach achieves greater accuracy of about 99.20% for EfficientNet. Meanwhile, 94.75% of accuracy is obtained for the Residual autoencoder network, whereas, the traditional methods like deep CNN-based autoencoder achieve 92.77%, which is lower than that of the proposed approach. Similarly, lower accuracy is achieved for various other existing approaches like Triplet network, CNN classifier, and various others. Thus, it is clear that the proposed work outperforms the state-of-the-art methods.

From Table 4, it is inferred that the proposed approach required only a smaller number of fine tunes than various other conventional methods that are used on datasets like FF, FF++, and DFD. Fine-tuning is nothing but a procedure in which the model that is already trained for a particular task is again tuned to complete a different related task. Similarly, the scratch of the proposed approach is also lower (52.82) when compared to that of the traditional methods. Hence, it is

Table 2 Performance analysis of the proposed model trained on Face Forensic++ dataset

Techniques	Metrics value (%)	
	Accuracy	AUC
Proposed EfficientNet	85.84	72.17
EfficientNet and LSTM	81.23	65.83
ResNetV2	76.04	59.42
VGG16 Inception	69.63	50.48

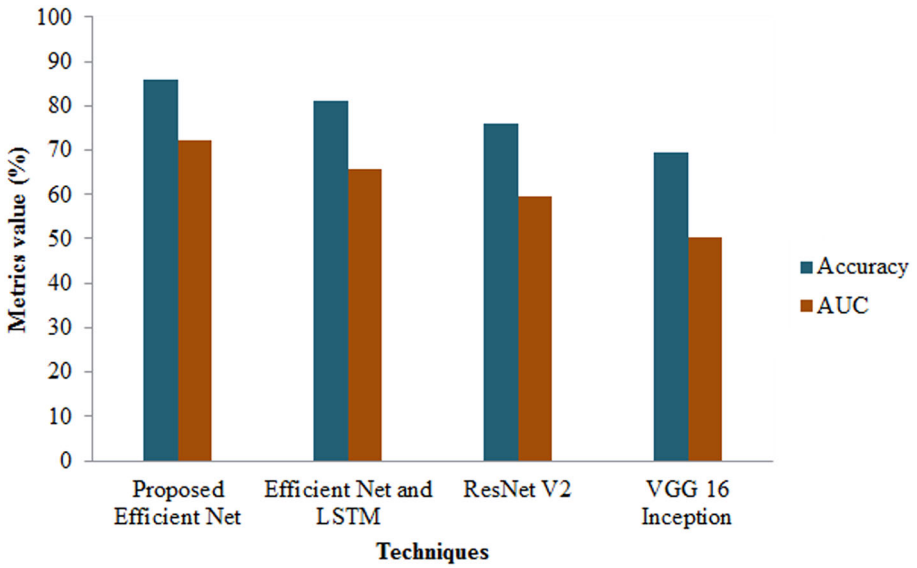

Fig. 4 Comparative analysis of the proposed model trained on Face Forensic++ dataset

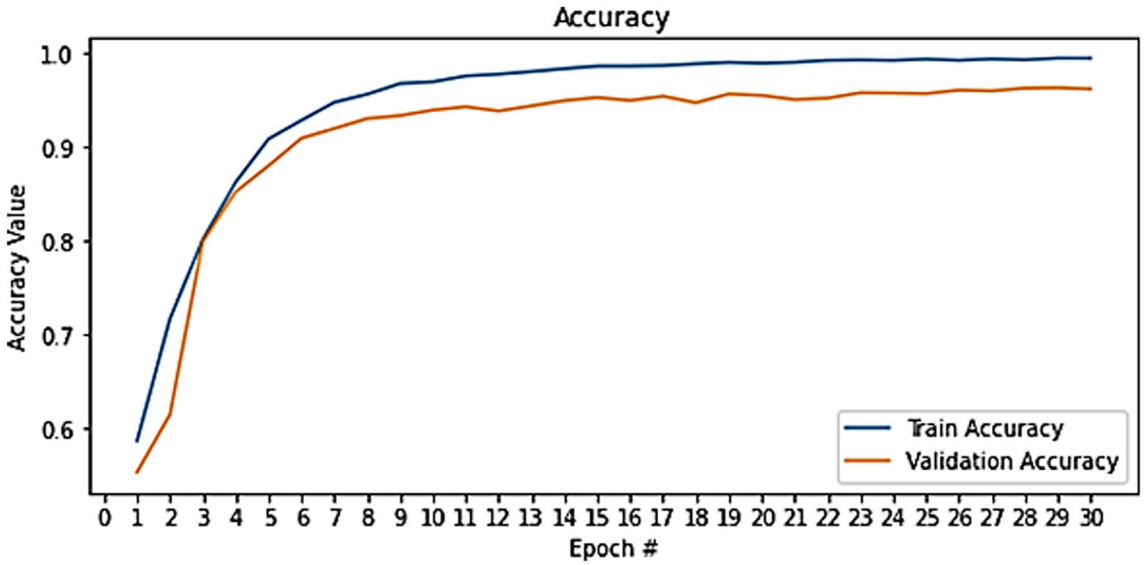

Fig. 5 Accuracy vs. Epochs for models trained on DFDC and FF++ datasets

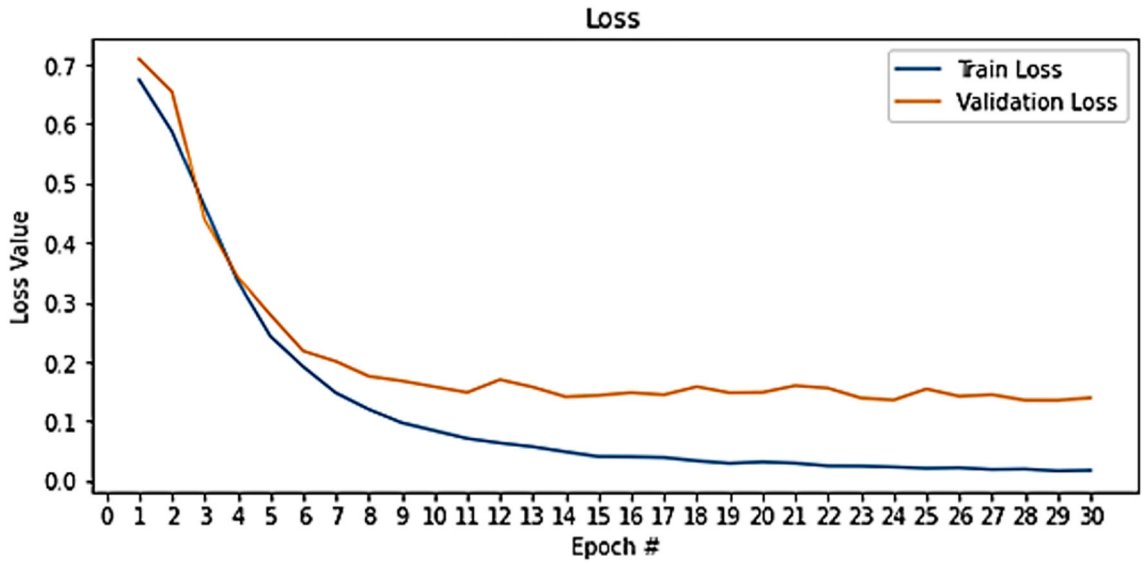

Fig. 6 Loss vs. Epochs for models trained on DFDC and FF++ datasets

clear that the proposed approach provides efficient detection of fake video in the network.

5.2 Performance evaluation of the proposed residual autoencoder network

The autoencoder model is trained separately from scratch for Face Forensics, Face Forensics++ and DFD datasets. Further, the model's accuracy is tested for other datasets. The results indicate that the model is overfitting for the dataset on which it is trained on. Hence the network is fine-tuned using dataset of different manipulation techniques and it is observed that transfer learning has boosted model's accuracy. The testing accuracy for all datasets is shown in Tables 5, 6 and 7.

The model trained using normal input images from FF++ datasets achieves an accuracy of 92.61% (Table 6) while the model trained using residual images achieves 94.75% accuracy (Table 7). The model trained using residual images shows an increase in accuracy as seen in Tables 6 and 7. It can be inferred that residual images help better to classify altered and original images.

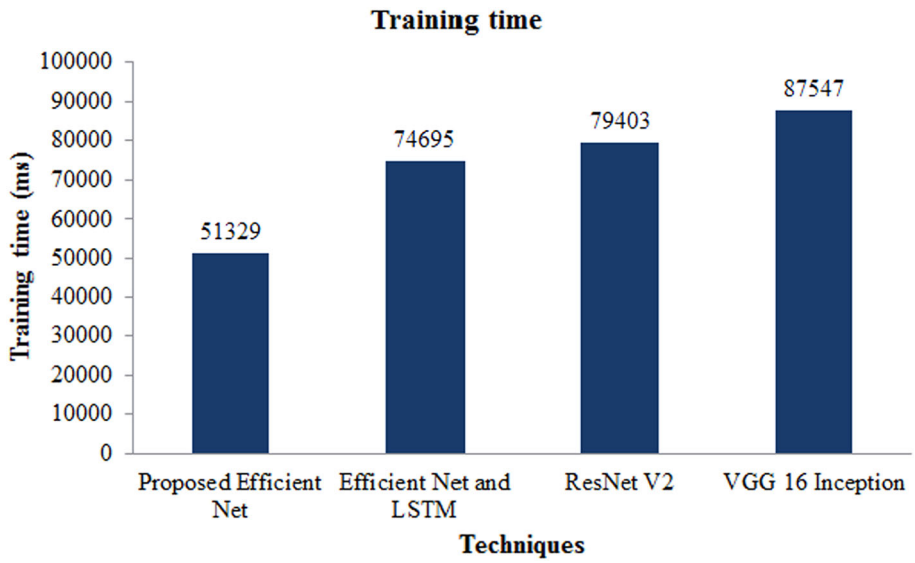

Fig. 7 Comparative analysis of proposed EfficientNet in terms of training time

Table 3 Comparison of the proposed method with the existing methods

PaperNo	Dataset	Method	Accuracy
Agarwal et al. [8]	DFDC + Custom	CNN + LSTM	97.60
MingXing et al. [20]	Face Forensic	Deep CNN-based autoencoder	92.77
Amerini et al. [22]	Face Forensics++	Triplet Network	86.74
Fei. J et al. [23]	Face Forensics++	CNN Classifier(VGG16)	81.61
Montserrat et al. [24]	Face Forensics++	Modified InceptionV3	98.99
Dufour [25]	DFDC	CNN + GRU	92.61
Proposed System	**Face Forensics++**	**Proposed CNN-based Method (EfficientNet)**	**99.20**
Proposed System	**Face Forensics++**	**Proposed Method (Residual Autoencoder Network)**	**94.75**

5.3 Performance evaluation of the proposed approach based on precision, recall and f-measure

From Fig. 8, it is clear that the presented approach obtains higher precision at the rate of 0.9% whereas the existing methods like CNN, RNN, and LSTM achieve 0.8%, 0.78%, and 0.71%. Similarly, the recall and f-measure value of the proposed technique lies in the range of 0.957% to 0.962% respectively. But, the existing approaches have lower (0.72%) recall and (0.74%) f-measure. Thus, it is clear that the proposed approach provides efficient accurate detection of fake videos.

Table 4 Results for model trained using Face Forensics dataset

Face Forensics Dataset		
Datasets	Scratch	Fine-tune
FF	93.85	97.07
FF++	54.83	69.64
DFD	52.82	57.02

6 Discussion

From the above analysis, it is understood that the proposed EfficientNet achieves better performance as compared to the existing methodologies. Most of the existing techniques work well for a small dataset, but, it shows performance degradation in large datasets. Thus, the proposed method mitigates this flaw and shows very good performance even for large datasets. Furthermore, most of the existing works require a huge time to train the data. But the proposed method trains the data with limited time and cost. Hence, it is concluded that the proposed method outperforms the other state of art methods.

Table 5 Results for model trained using Face Forensics++ dataset

Face Forensics++ Dataset		
Datasets	Scratch	Fine-tune
FF	50.43	79.50
FF++	92.61	96.61
DFD	68.61	80.73

Table 6 Results for model trained using DFD dataset

DFD Dataset		
Datasets	Scratch	Fine-tune
FF	45.91	76.41
FF++	72.11	75.90
DFD	89.47	90.86

Table 7 Results for model trained using residual images

Datasets	Accuracy
FF	54.21
FF++	94.75
DFDC	74.11

7 Conclusion

With the realistic nature of deep fake, it is becoming a major threat in our society, business and politics. Therefore, it is very essential to tackle the problem of deep fakes on supremacy. Authors have proposed deep fake detection models using different features like audio, visual artifacts, frequency spectrum, statistical, audio-visual mismatch etc. but the existing models are not sufficient for reliable fake detection. The existing techniques to detect deep fakes are not robust for all kind of manipulation techniques as new manipulation types keep evolving. Model trained on specific manipulation technique does not perform well for unseen manipulations and datasets. Their performance for unseen manipulations and datasets drops drastically which makes them unfit for practical applications. In this paper, the problem of generalizability is analysed in order to make the network robust to all kind of attacks. Paper highlights on improvising the accuracy of deep fake detection models using transfer learning approach. Fine-tuned models are able to provide better accuracy as compared to model strained from scratch. The results also show that residual image inputs increases the model's accuracies. Efficient Net based model trained on datasets like DFDC and Face Forensics++ achieves AUC score of 94% on DFDC and 98% on Face Forensics++.

The future work focuses mainly on exploring more deep learning models with reduced parameters for the task of video based Deepfake detection. The task of deep fake detection is highly affected by the variety of data used to train the models. Several benchmarks datasets can be explored to train the models with variations. A custom data set can be synthesized with combination of variations to test the model for detecting unseen attacks. The proposed system focuses on detecting a manipulated face of person in a frame, but there are other evolving deep fakes with several manipulated faces in a frame. Future work proposes extension of the proposed systems to detect such deep fakes with group of people consisting of a single fake face or multiple fake faces in a video. The facial expressions of faces can be recognized [26] and

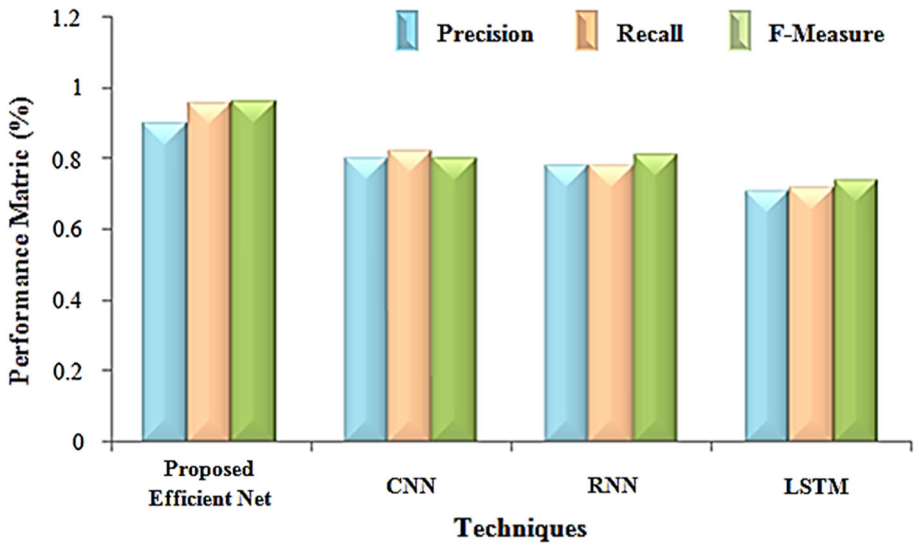

Fig. 8 Comparative analysis of proposed Efficient Net based on precision, recall and f-measure

differentiated between fake and real by CNN. A multimodal deep fake detection system can be designed by incorporating audio deep fake detection in the proposed video deep fake detection.

Acknowledgements The authors would like to thank CoE-CNDS Lab from VJTI, Mumbai to provide NVIDIA's DGX-1 required for training the models. The authors are grateful to the entire team lab to provide support during COVID-19 pandemic.

Data availability The datasets generated analyzed during the study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest Nil.

References

1. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer C (2019) The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv:1910:08854
2. Rössler, A; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M.: FaceForensics: A large-scale video dataset for forgery detection in human faces (2018)
3. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M.: Face- Forensics++: learning to detect manipulated facial images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (2019). https://doi.org/10.1109/ICCV.2019.00009
4. N. Dufour, A. Gully, P. Karlsson, A. V. Vorbyov, T. Leung, J. Childs, and C. Bregler (2019) Deepfakes detection dataset by google & jigsaw
5. Li, Haodong; Li, Bin; Tan, Shunquan; Huang, Jiwu: Identification of deep network generated images using disparities in color components. Signal Process. (2020). https://doi.org/10.1186/s13635-020-00109-8
6. Ding, X., Raziei, Z., Larson, E.C. et al. (2020)
7. Hsu, C.-C.; Zhuang, Y-Xiu.; Lee, C.-Y.: Deep fake image detection based on pairwise learning. Applied Sciences **10**(1), 370 (2020). https://doi.org/10.3390/app10010370
8. Hashmi, M.F.; Ashish, B.K.K.; Keskar, A.G.; Bokde, N.D.; Yoon, J.H.; Geem, Z.W.: An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture. IEEE Access **8**, 101293–101308 (2020)
9. Agarwal S, Farid H, Fried O and. Agrawala M (2020) Detecting Deep-Fake Videos from Phoneme- Viseme Mismatches. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA
10. Minh Dang, L.; Hassan, S. I.; Im, S.; Moon, H.: Face image manipulation detection based on a convolutional neural network. Expert Syst. Appl. **129**, 156–168 (2019). https://doi.org/10.1016/j.eswa.2019.04.005
11. Matern F, Riess C. Stamminger M (2019) Exploiting Visual Artifacts to Expose Deep fakes and Face Manipulations.2019 In: IEEE Winter Applications of Computer Vision Workshops (WACVW),Waikoloa Village, HI, USA
12. Sabir, E.; Cheng, J.; Jaiswal, A.; Almageed, W. A.; Masi, I.; Natarajan, Prem: Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. IEEE Conf. Comput. Vision Pattern Recogn. **3**, 80–87 (2019)
13. Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian (2018) Not made for eachother– Audio-Visual Dissonance-based Deepfake Detection and Localization. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05
14. Y. Li, M. Chang and S. Lyu,"In Ictu Oculi (2018) Exposing AI Created Fake Videos by Detecting Eye Blinking.In: IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong,Hong Kong
15. Li Y,.Lyu S (2018) Exposing Deep Fake Videos By Detecting Face Warping Artifacts. In: IEEE Conference Computer. Vision Pattern Recognition.
16. Torfi, A.; Iranmanesh, S.M.; Nasrabadi, N.; Dawson, J.: 3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition. IEEE Access **5**, 22081–22091 (2017)
17. Matthews, T.F.; Cootes, J.A.; Bangham, S.C.; Harvey, R.: Extraction of visual features for lip reading. IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 198–213 (2002)
18. H. Li, H. Chen, B. Li and S. Tan (2018) Can Forensic Detectors Identify GAN Generated Images. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPAASC),
19. Honolulu, HI, USA, HSheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, Alexei A. Efros (2019) CNN- generated images are surprisingly easy to spot... for now. In: IEEE Conference on Computer Vision and Pattern Recognition
20. HuyH.Nguyen, Fuming Fang, Junichi Yamagishi, Isao Echizen (2019) Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. In: IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS).
21. MingxingTan,QuocV.Le(2019)EfficientNet:Re- thinking Model Scaling for Convolutional Neural Net- works. Cornell University
22. M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, NA. Kumar, A. Bhavsar and R. Verma (2020) Detecting Deepfakes with Metric Learning.In: 8th International Workshop on Biometrics and Forensics (IWBF), Porto, Portugal
23. I. Amerini, L. Galteri, R. Caldelliand A. DelBimbo (2019) Deepfake Video Detection through Optical Flow Based CNN In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South)
24. Fei, J.; Xia, Z.; Peipeng, Yu.; Xiao, F.: Exposing AI-generated videos with motion magnification. Multimedia Tools Appl. **80**(20), 30789–30802 (2020). https://doi.org/10.1007/s11042-020-09147-3
25. D. M. Montserrat et al., (2020) Deepfakes Detection with Automatic Face Weighting In: IEEECVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA
26. Umer, S.; Rout, R.K.; Pero, C.etal.: Facial expression recognition with trade-offs between data augmentation and deep learning features. J Ambient Intell Human Comput (2021). https://doi.org/10.1007/s12652-020-02845-8