

## A Novel Method of Deepfake Detection

Rutuja Hande<sup>1,\*</sup>, Sneha Goon<sup>1,\*\*</sup>, Aaditi Gondhali<sup>1,\*\*\*</sup>, and Navin Singhaniya<sup>1,\*\*\*\*</sup>

<sup>1</sup>Department of Electronic Engineering, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai

**Abstract.** Deep-Fake is a novel artificial media technology that uses the likeness of someone else to replace people in existing photographs and films. Deep Learning, as the name implies, is a type of Artificial Intelligence that is used to create it. It is critical to develop counter attacking approaches for detecting fraudulent data. This research examines the Deep-Fake technology in depth. The Deep-Fake Detection discussed here is based on current datasets, such as the Deep-Fake Detection Challenge (DFDC) and Google's Deep-Fake Detection dataset (DFD). The creation of a bespoke dataset from high-quality Deep-Fakes was utilised to test models. The results of both with and without Transfer Learning were analysed. Finally, the trained models were used to spot well-known deep-fakes of former US President Barack Obama and well-known actor Tom Cruise. A comparison study was performed on all three models. The findings show that the detection are generally domain-specific tasks, however that using Transfer Learning considerably improves the model performance parameters, whereas convolutional RNN gives sequence detection advantage.

### 1 Introduction

The internet and social media platforms have brought people from all corners of the globe closer together. In a world where recording and sharing are an inevitable part of everyday life, it's critical to be aware of the far-reaching consequences that Deep-Fakes can have [1]. However, This is no longer the case, thanks to the appearance of Deep Fake Video. Technology has the ability to convince others that something is real when it isn't." Deep-Fakes is a combination of the words "deep" and "fake." "Deep learning" and "fake" are two words that come to mind.

In this technology, autoencoders and GANs are utilised to create visual and auditory data. The resulting information is rife with deception [2]. Autoencoders are neural networks that retrieve and change the original image dimension by extracting crucial facial traits. Traditionally in the latent space The latent space is used to depict data analysis in a more realistic way. In a transparent way The de-noising process will overlay the original image's features [3]. This image was coded by a coder who had been specifically trained for it. Incorporation of a Generative The addition of an adversarial network to the decoder side helped to increase the technique's generation.

If unmanaged, this cutting-edge technology can be disastrous, as the era of believing what one sees is over. However, malevolent Deep-Fakes dominate the productive applications of Deep-Fakes, such as film dubbing, special effects, and instructional objectives [5]. Information



**Figure 1.** Frames from Deep-Fake video of former US President Donald Trump

Warfare, Celebrity Defamation, Political Propaganda, Scams and Fraud, Harassment, and Blackmailing are all examples of how this weapon of mass misinformation can be utilised. With the enhanced quality of Deep-Fakes created with deep learning algorithms like GANs and Autoencoders, it's getting increasingly difficult to tell them apart with human eyes. The project's vision is to develop Deep-Fake video detection frameworks that are trained on the most recent video datasets and then assess the results. Convolution Neural Network architecture was used to detect deep fakes [9]. A hybrid approach of Convolutional Recurrent Neural Network is used to diverse video datasets, taking into account the added temporal nature of video. LSTM or GRU can be used as the RNN's later extension network. Comparative analysis of the results is performed on some of the most recent DeepFake datasets, providing insight into the types of data

\*e-mail: rutujahande2000@gmail.com

\*\*e-mail: snehagoon2000@gmail.com

\*\*\*e-mail: aaditi.gondhali28@gmail.com

\*\*\*\*e-mail: navin.singhaniya@rait.ac.in

dealt with by each of them on all three model architectures as discussed, and is also represented using tabular and graphical representation for better visual understanding [13].

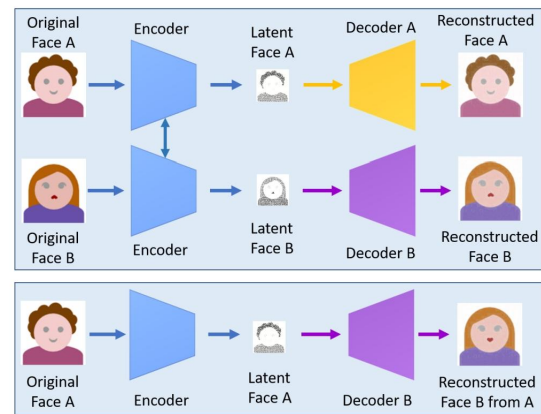
## 2 IMPLEMENTATION

### APPROACHES BASED ON CNN+RNN NETWORKS

The extraction of features from video frames is done using a Convolutional Neural Network. RNN is trained on these features in order to classify videos into one of two categories: fake or authentic. Another model based on the ConvLSTM hybrid architecture [2] examines the minute visual markers on the faces using CNN and then evaluates the visual data using an RNN network based on the features. The widespread availability of resources such as GPUs has resulted in the widespread distribution of these videos. RNN is well-known for its ability to analyse sequential data. We already know that the temporal information extracted from video frames is sequential. As a result, RNN can be used to process it. The work presented by Sabir et al. [3] used the best model for recognising facial alterations, which uses the RNN network. The FaceForensics++ dataset was used for testing, and efficiency was enhanced by 4.55 percent above the prior state-of-the-art. Some approaches focus on the faces in the films to reliably detect modifications, extracting visual and temporal information from the faces. These methods employ a recurrent neural network-based architecture to investigate temporal trends. D. M. Montserrat et al. [4] explored one such strategy. They compared the results with previous methodologies using the DFDC dataset. The proposed method produces more accurate findings than existing methods. MTCNN is the CNN model employed, and the model is used for temporal feature extraction. The proposed model has a validation accuracy of 92.61 percent and a test accuracy of 91.88 percent.

### APPROACHES BASED ONLY CNN NETWORKS

Face warping is used in Deep-Fake generation, as are numerous morphing techniques like as scaling, translation, and affine deformations. Deep-Fake generation's warping method causes resolution discrepancy [7] between the warped face area and the surrounding back-ground area. Visual artefacts [8] are created in the video frames as a result of this. This type of technique takes advantage of the errors that occur throughout the generation process. Convolutional Neural Network models are used in these methods to detect the artefacts. Convolutional Neural Networks could be quite effective for computer-assisted detection difficulties. H.-C. Shin et al. present a paper [9] that uses deep CNN factors to solve Computer-Aided Detection (CADe) problems. Another CNN-based effort on image modification detection [10]



**Figure 2.** deepfaken creation model using two encoder-decoder pair [3]

use a high pass filter in conjunction with a CNN network to detect the image's hidden features.

### 2.1 ALGORITHM

For the detection of Deep-Fake films, a variety of existing models have been applied. A comparison of CNN-based video deep fake detection with CNN followed by RNN-based Video DeepFake Detection is discussed in this paper. The performance of two RNN variants, LSTM and GRU, on temporal feature exploitation is compared. The algorithm of the method presented in this paper is discussed in this part. The major goal of the system under discussion is to determine whether the video input is authentic or Deep-Fake. This goal can be accomplished in three different ways. Three different models are compared based on their performance for Video Deep-Fake Detection. The following is the algorithm for these processes:

**STEP 1:** Frames from the input video are extracted. The input video is preprocessed and turned into frames in this step. The CNN architecture is fed a predetermined amount of frames as an input.

**STEP 2:** Each frame's features are extracted. For an unseen test sequence input, CNN extracts a collection of facial traits for each frame. The feature set is concatenated when this operation is performed on a preset number of frames at the same time.

**STEP 3:** Analyze Temporal Sequences[1] This stage is different in each of the three models covered in the paper. To acquire a single probability value for the CNN model, the concatenated features are input to dense, followed by a global average pooling layer. Concatenated features are given to LSTM and GRU layers for temporal

Dataset	realvideos	fakevideos	totalvideos	rightcleared	agreeingsubject	total subjects	synthesismethods
UADFV	49	49	98	NO	0	49	1
DeepFakeTIMIT	640	320	960	NO	0	32	2
FF++	1000	4000	5000	NO	0	N/A	4
CelebDF	590	5639	6229	NO	0	59	1
GoogleDFD	363	3000	3363	YES	28	28	5
DeeperForensics &50000	10000	60000	YES	100	100	1	
DFDC	23654	104500	128154	YES	960	960	8
KoDF	62166	175776	237942	YES	403	403	6
FkeAVCeleb	500	19500	20000	NO	0	500	4

**Table 1.** Quantitative comparison of fakeAVCeleb to existing publically available deepfake dataset

feature analysis in CNN followed by LSTM and GRU models, and a single probability value is obtained.

STEP 4: Classification of video as real video or Deep-Fake video From the probability values obtained in step 3, the unseen test input video is classified as either manipulated i.e. DeepFake video, or non-manipulated video i.e real video on the basis of decided threshold

## 2.2 Mathematical Modelling

Each layer of CNN can be as follows:

- Convolutional Layer
- Pooling Layer
- Fully connected layer
- Convolutional Layer

Convolutional product is applied to this layer using many filters followed by an activation function .

- Pooling Layer The pooling layer’s goal is to down sample the input’s features without reducing the.the following notations are taken into account:
- Fully Connected Layer It is a finite number of neurons. It takes input as a vector as well as returns a vector.

### Maths Behind LSTM-

$$Ct = Ct + (ItC' t)$$

Calculating output

$$Ht = \tanh(Ct)$$

### Maths Behind GRU-

Consider time step t and input is a minibatch  $XtRnxd$  , where n is number of examples and d is number of inputs. Hidden state of earlier time step is  $Ht1Rnxh$ . From this we can calculate reset gate  $RtRnxd$  and update gate  $ZtRnxd$  as follows:

$$Rt = (XtWxr + Ht1Whr + br)$$

$$Zt = (XtWxz + Ht1Whz + bz)$$

Here W are the weights and b are the biases. Candidate Hidden State  $H\tilde{t}Rnxd$  state is given as:  
 $H\tilde{t} = \tanh(XtWxh + (RtJHt1)Whh + bh)J$  is the

Hadamard Product Operator.

### Proposed Model Architecture

More preciously, at the **1st layer**, we denote:

- INPUT:  $a^{[l-1]}$  with size  $(n_H^{[l-1]}, n_W^{[l-1]}, n_C^{[l-1]})$ ,  $a^{[o]}$  being the image in the input
- PADDING:  $p^{[l]}, Stride : s^{[l]}$
- NUMBER OF FILTER:  $n_C^{[l]}$  where each  $K^{(n)}$  has the dimension  $(f^{[l]}, f^{[l]}, n_C^{[l-1]})$
- BIAS OF nth CONVOLUTION:  $b_n^{[l]}$
- ACTIVATION FUNCTION:  $\psi^{[l]}$
- OUTPUT:  $a^{[l]}$  with size  $(n_H^{[l]}, n_W^{[l]}, n_C^{[l]})$

And we have

$$\forall n \in [1, 2, \dots, n_C^{[l]}]$$

$$\text{conv}(a^{[l-1]}, K^{(n)})_{x,y} =$$

$$\psi^l(\sum_{i=1}^{n_H^{[l-1]}} \sum_{j=1}^{n_W^{[l-1]}} \sum_{k=1}^{n_C^{[l-1]}} K_{i,j,k}^n a_{x+i,y+j-1,k}^{[l-1]} + b_n^l)$$

$$\text{dim}(\text{conv}(a^{[l-1]}, K^{(n)})) = (n_H^l, n_W^l)$$

Thus,

$$a^{[l]} = [\psi^{[l]}(\text{conv}(a^{[l-1]}, K^{[n]})), \psi^{[l]}(\text{conv}(a^{[l-1]}, K^{[2]})), \dots, \psi^{[l]}(\text{conv}(a^{[l-1]}, K^{[n_C^l]})]$$

$$\text{dim}(a^{[l]}) = (n^{[l]}H, n^{[l]}W, n_C^{[l]})$$

with,

$$n_{H/W}^{[l]} = \left\lceil \frac{n_{H/W}^{[l-1]}}{s^{[l]}} + 1 \right\rceil; s > 0$$

$$= n_{H/W}^{[l-1]} + 2p^l - f^l; s = 0$$

$$n_C^l = \text{number of filters}$$

The learned Parameters at the  $l^{th}$  layer are:

- FILTERS with  $(f^{[l]} \times f^{[l]} \times n_C^{[l-1]}) \times n_C^{[l]}$  Parameters
- BIAS with  $(1 \times 1 \times 1) \times n_C^l$  Parameters

### Pooling Layer Parameters

$$a_{x,y,z}^l = \text{pool}(a^{[l-1]})_{x,y,z} = \phi^l(a_{x+i-1,j-1,z}^{[l-1]}), i, j \in [1, 2, 3 \dots f^l]$$

$$\dim(a^l) = n^l H, n^l W, n_C^l$$

$$n_{H/W}^{[l]} = \left\lfloor \frac{n_{H/W}^{[l-1]}}{s^{[l]}} + 1 \right\rfloor; s > 0$$

$$= n_{H/W}^{[l-1]} + 2p^{[l]} - f^{[l]}; s = 0$$

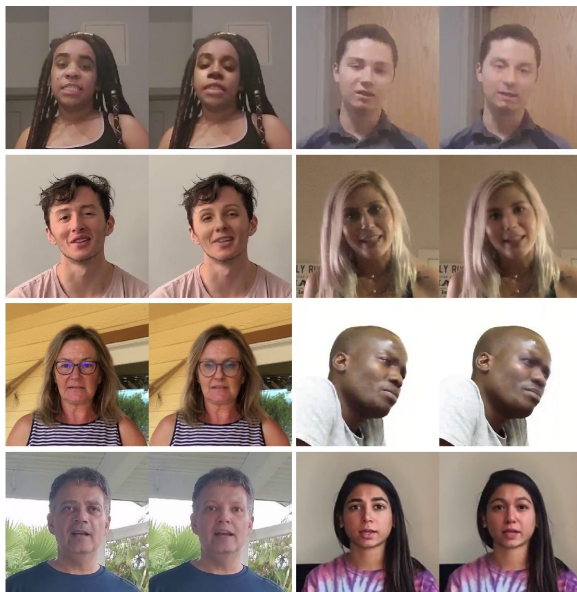
$$n_C^{[l]} = n_C^{[l-1]}$$

### Fully Connected Layer Parameters

- INPUT:  $a^{[l-1]}$  with size  $(n_H^{[l-1]}, n_W^{[l-1]}, n_C^{[l-1]})$ ,  $a^{[0]}$  being the image in the input
- PADDING:  $p^l$  rarely used, (stride):  $s^{[l]}$
- SIZE OF THE POOLING FILTER:  $f^{[l]}$
- POOLING FUNCTION:  $\text{phi}^{[l]}$
- OUTPUT:  $a^{[l]}$  with size  $(n_H^{[l-1]}, n_W^{[l-1]}, n_C^{[l-1]})$

### 2.3 DATA SET

Various already existing as well as unique datasets were used to train and test the models. In this section, we'll go through a quick overview of each dataset we'll be using. The datasets utilised for Video Deep-Fake Detection are as follows:



**Figure 3.** Deep-Fake Detection Challenge Dataset [4]

**1. Deep-Fake Detection Dataset (DFD):** This is the Google/Jigsaw Deep-Fake Detection Dataset. It has 3068

Deep-Fake videos, which were created from 363 original videos of 28 persons of various genders, ages, and cultural backgrounds.

### 2. Deep-Fake Detection Challenge Dataset (DFDC):

This is the Deep-Fake Detection Challenge Dataset from Facebook. There are two versions of it. 5K films made by two facial modification algorithms are included in the preview edition. There are 124k videos made by eight facial modification algorithms in the entire dataset version.

**3. Custom Dataset:** A custom dataset is generated by mixing 936 from existing Video Deep-Fake Detection Datasets, such as DFDC and DFD datasets, to test the generalizability of an algorithm.

## 3 PERFORMANCE PARAMETER

Performance Parameter To evaluate the outputs of three deep neural networks: CNN, CNN-LSTM, and CNN-GRU, the following metric parameters are used to compare their performance:

- Accuracy:- The most common comparative metric is accuracy. The ratio of the total number of correctly classified examples to the total number of examples classified is known as the correct classification rate. It is the most commonly used metric for comparing models.
- Precision:- The ratio of successfully categorised positive examples divided by the total number of anticipated positive examples is the precision value. A high precision number indicates that an example classified as positive is, in fact, positive.
- recall:- The ratio of the total number of correctly categorised positive examples divided by the total number of positive examples is known as recall. A high recall rate shows that the class has been appropriately identified.
- F1 score:- The combined effect of precision and recall value is represented by the F1 score. It's the sum of recall and precision in a harmonic form. summarise the results of the CNN model, CNN-LSTM model, and CNN-GRU model trained and tested on the DFD dataset, DFDC dataset, and Custom dataset, respectively. The variation of the metric parameters described above is obtained over a number of epochs. It's also worth noting that the outcomes of the models with and without finetuning are nearly identical.

## 4 OUTPUTS AND RESULT ANALYSIS

In this sections results of each model based on their performance on various datasets, various testing parameters, and training parameters is discussed by graphical means. This section also includes some popular video Deep-Fake detection tested using the models described in the report and

comparative analysis between three models: CNN, CNN-LSTM, CNN-GRU. The results of CNN model, CNN-LSTM model and CNN-GRU model trained and tested on DFD dataset , DFDC dataset and Custom dataset. The variation of above mentioned metric parameters are obtained over number of epochs. And it is observed that with and without finetuning the models results are similar.

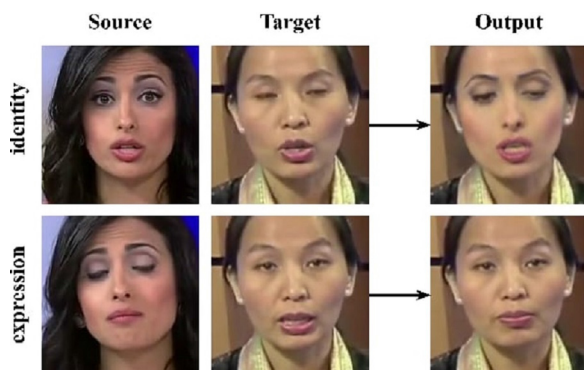


Figure 4. Output

• Deepfake Video ft. Former President Barack Obama



Figure 5. Output

## 5 CONCLUSION

Seeing is Believing’ was formerly thought to be true. It’s no longer true, given the emergence of Deep-Fake content. Deep-Fake has begun to erode people’s faith in media information. Deep-fake films have the potential to wreak havoc on politics, slander individuals, disseminate hate speech, and the list goes on. As a result of the spread of this, detecting it effectively becomes critical.

This paper describes a method for detecting Deep-Fake videos that employs Transfer Learning. By leveraging pre-trained networks, the work provides valuable insights into the detection (with comparative aspect) of AI-generated bogus videos.

Frame Extraction and detection are used to successfully achieve Deep-Fake detection of videos in this work. Frame detection involves obtaining a sequence of frames, which are then analysed for temporal feature changes using CNN, LSTM, and GRU. Using the CNN model, we retrieved face features from numerous consecutive frames and passed them to the network. The hybrid framework of

• Deepfake Video ft. Tom Cruise



Figure 6. Output

CNN layer followed by LSTM proves to be relevant for the task of Video Deep-Fake Detection, according to the findings obtained by performing tests on different datasets. In a Deep-Fake Video, the CNN-LSTM pipeline works in the domain of temporal discontinuities in successive frames. The CNN model is followed by the CNN-LSTM model, which also does well in detecting Fake frames since the detection pipeline collects frames from videos and feeds them into the model.

## 6 FUTURE WORK

According to the literature review, Deep-Fake detection is one of the most popular research topics in the AI sector. To detect alterations in the Deep-Fake media, a variety of methods based on neural networks and biological signals have been used.

So far, the research has concentrated on the visual and audio components separately. So, in order to do the Deep-Fake Video as well as Audio Detection at the same time on the same video, future work will entail implementing Deep Temporal Convolutional Model (TCN) for deeper analysis and testing. Also, rather than using interpretability methods to those models as the methodology utilised, work on developing models that operate as a generalising architectural foundation for the majority of sorts of Deep-Fake techniques.

Although much research has been done on the development and detection of Deep-Fake, not enough has been done on the reversal of those manipulations.

As a result, a network may be created that can undo the Deep-Fake video or image, allowing us to reconstruct the original video or image from the Deep-Fake one. Seeing the consequences that Deep-Fakes have on human life it is a good lesson to double check the content before sharing it on any internet platforms

## References

- [1] Mitra, Alakananda, Saraju P. Mohanty, Peter Corcoran, and Elias Kougianos. "A novel machine learning based method for deepfake video detection in social media." In 2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS), pp. 91-96. IEEE, 2020.
- [2] Mitra, Alakananda, Saraju P. Mohanty, Peter Corcoran, and Elias Kougianos. "A machine learning based approach for deepfake detection in social media through key video frame extraction." *SN Computer Science* 2, no. 2 (2021): 1-18.
- [3] Güera, David, and Edward J. Delp. "Deepfake video detection using recurrent neural networks." In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), pp. 1-6. IEEE, 2018.
- [4] Zhao, Tianchen, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. "Learning self-consistency for deepfake detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15023-15033. 2021.
- [5] Zhao, Hanqing, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. "Multi-attentional deepfake detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2185-2194. 2021.
- [6] Khalid, Hasam, Shahroz Tariq, Minha Kim, and Simon S. Woo. "FakeAVCeleb: a novel audio-video multimodal deepfake dataset." arXiv preprint arXiv:2108.05080 (2021).
- [7] Jafar, Mousa Tayseer, Mohammad Ababneh, Mohammad Al-Zoube, and Ammar Elhassan. "Forensics and analysis of deepfake videos." In 2020 11th international conference on information and communication systems (ICICS), pp. 053-058. IEEE, 2020.
- [8] Rana, Md Shohel, and Andrew H. Sung. "Deepfakestack: A deep ensemble-based learning technique for deepfake detection." In 2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom), pp. 70-75. IEEE, 2020.
- [9] Deshmukh, Anushree, and Sunil Wankhade. "Deepfake detection by exposing ai-generated fake face video." In Proceedings of Integrated Intelligence Enable Networks and Computing, pp. 673-679. Springer, Singapore, 2021.
- [10] Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. "Deepfakes and beyond: A survey of face manipulation and fake detection." *Information Fusion* 64 (2020): 131-148.
- [11] Cozzolino, Davide, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. "Id-reveal: Identity-aware deepfake video detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15108-15117. 2021.
- [12] Pu, Jiameng, Neal Mangaokar, Lauren Kelly, Parantapa Bhattacharya, Kavya Sundaram, Mobin Javed, Bolun Wang, and Bimal Viswanath. "Deepfake videos in the wild: Analysis and detection." In Proceedings of the Web Conference 2021, pp. 981-992. 2021.
- [13] Wodajo, Deressa, and Solomon Atnafu. "Deepfake video detection using convolutional vision transformer." arXiv preprint arXiv:2102.11126 (2021).
- [14] Lu, Yuhang, and Touradj Ebrahimi. "A New Approach to Improve Learning-based Deepfake Detection in Realistic Conditions." arXiv preprint arXiv:2203.11807 (2022).