Deepfake Video Detection using Neural Networks

Nimitt Patel[1], *Niket* Jethwa[2], *Chirag* Mali[3], *Jyoti* Deone[4]
[1,2,3,4] Department of Information Technology Engineering
[1,2,3,4] Ramrao Adik Institute of Technology University, DY Patil Deemed to be University, Nerul, Maharashtra, India

Abstract—In today's era, software tools based on deep learning have made the people work easier to make credible faces exchanges in video with little signs of manipulation, nicknamed "DeepFake" videos. Manipulation in digital media has been performed for decades through the appropriate use of visual effects; nevertheless, current breakthroughs occurred in deep learning have resulted in a significant rise to gain reality of fake material or contents using the simple ways. This are Artifical Intelligence-generated media (known as DF). Using tools of artificial intelligence to create the DF is an easy task. However, detecting these DF poses a significant barrier. Because it is difficult to teach the algorithm to detect the DF. Using Convolutional Neural Networks and Recurrent Neural Networks, we have made progress in detecting the DF. The system employs a Convolutional Neural network (CNN) on frame level to extract features. These observations are noted and this can train a Recurrent Neural Network (RNN), which has the ability to learn and classify whether or not a video has been tampered with and identified the temporal irregularities in the frame introduced by DF tools. We demonstrate how utilizing a simple architecture, our system may get competitive outcomes in this job.

I. INTRODUCTION

The growing era of mobile technology and integration of cameras, as well as the expanding reach of social media and sharing media portals, has made the creation and dissemination of digital video easier than before. Lacking in the advanced tools and high demand to expertise the time-consuming steps which are difficult and steps involved to limit the false videos and degree of realism until recently. However, the required time to create and manipulate videos has been reduced in the past years, this all is possible because of large amounts of training data and computing power, majorly the advancements in computer vision techniques and machine learning that replaces the requirement of manual editing. [1]

Tools like the Adobe Photoshop are used for video editing, but editing videos by replacing the faces is tedious task for this software, like if we want to process 20 second video with 25 frames per second, then it will edit about 500 images. So, software like this cannot edit this large number of images. [2]

Nowadays, any small video of any person or identity of a person can be forged very easily by replacing the facial image. [4] A fully deepfake audio-video of any person can be created by the techniques developed by Suwajanakorn and othes. [4]

A lot of attention has been attracted recently by the new vein of fake video generation using AI-based technology for its generation. It takes an input video of a particular individual and provides an output video with the individual's face replaced with another person's and the result is provided. [6]

Deep neural networks developed and trained on face images to automatically map and detect expressions of facials from the source to target which act as a backbone for DeepFake video generation. A high level of realism is achieved with effective post-processing. [1]

The importance of DF detection in such a situation cannot be overstated. As a result, we present a novel deep learning-based strategy for distinguishing false videos generated by AI technology from actual(real) videos. It's critical to have technologies that can detect fake videos so that they can be tracked down and avoided from getting viral over the internet. An example of deepfake is show in Figure 1.

Fig 1. A deepfake manipulate images examples [14]

It is critical to comprehend how the Generative Adversarial Network (GAN) generates the DF in order to detect it. GAN takes a video and extracts an image of a person (target) as input and provides a video with the face of target being replace with another person's face (source). Deep learning alongside neural networks being trained on the face cropped photos and target videos provides the backbone of DF, which automatically transfers the source's faces and facial emotions to the target. [1]

The produced movies can achieve a high level of realism with suitable post-processing. The GAN performs the function of breaking the videos down into frames and replacing each frame with input image. It goes on to rebuild the video.

Autoencoders are commonly used to do this. We provide a new deep learning-based strategy for distinguishing DF videos from actual real-world videos. The solution is based on the same mechanism as GAN's DF creation. [7] The approach is based on DF video attributes; because of production time constraints and

computational resources, the DF algorithm only synthesizes face pictures of limited size and must undergo the step of affinal warping to fit and save the source's face configuration. Due to the inconsistency in resolution between the surrounding context warped face area, this warping leaves some noticeable artifacts in the output deep fake video. [1]

By splitting the video into frames and comparing the created face areas and their surrounding regions, our approach detects such artifacts. Using a LSTM along with RNN to capture the inconsistencies between frames produced by GAN during the process of DF reconstruction which is temporal, and getting the features with a ResNext Convolutional Neural Network. [8]

The process of training the ResNext CNN model is simplified by making models of inconsistency in affine face wrapping. The GAN consist of a discriminator which is simply a classifier. It is used to differentiate between the real data from data generated. [15-17]

A simple GAN architecture diagram is shown in Figure 2, which gives a clear idea about how it works, and processes the image as fake or real.

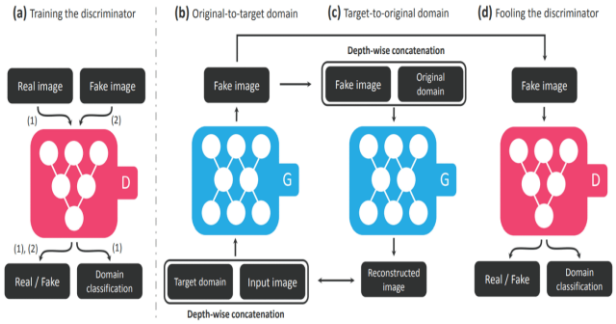

Fig. 2. GAN's General Architecture

II. LITERATURE SURVEY

Deep fake video's fast development and illicit use pose a serious danger to democracy, justice, and public trust. As a result, the demand for fraudulent video detection, analysis, and intervention has increased. Following are some of the relevant words in deep fake detection:

The method employed in ExposingDF Videos by Detecting Face Warping Artifacts[1] was to compare the surrounding regions of the face and the relative face area with artifact detection using Convolutional Neural Network. This work consists of two Face Artifacts.

Such an idea was based on the observation that images with limited resolutions must be altered further to match the source face in the approach of the present algorithm being applied over the video.

The paper Exposing AI-Created False Videos by Detecting Eye Blinking[2] explains a unique method of exposing deepfake videos which are created using deep neural network models. This approach depends on the identification of blinking of eyes in the video as it's a physiological signal which is difficult to present in bogus videos.

The method works on the eye blinking datasets and provides promising results when it comes to detecting videos created with the Deep Neural Network-based program DF.

Their strategy relies solely on the lack of blinking as a detection clue. However, additional factors such as teeth enchantment, wrinkles on the face, and so on must be considered when detecting a thorough fake. All these parameters are considered in our project.

Detecting forged images and videos [3] with capsule networks is a method that employs a capsule network to detect forged, modified photos and videos in a variety of circumstances, such as computer-generated video detection and replay attack detection. Figure 3 shows the total number of papers published from the year 2016-2021.

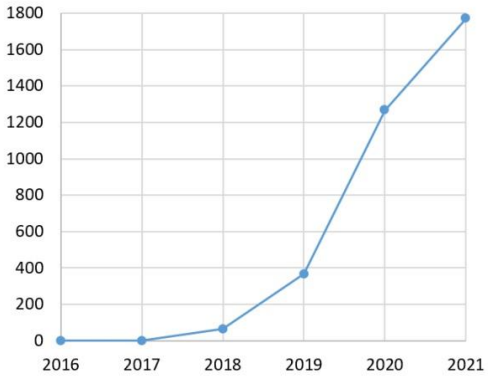

Fig 3. Graph showing the count of papers published from 2016-2021 based on deepfake project [18-20]

Our approach is designed to be trained on datasets that are both noiseless and real-time. Comparison between the different papers is shown in Table 1 based on the key parameters that distinguish each project.

Table 1. Comparison Between Different Projects

Parameters	**Exposing DF Videos by Face Warping Artifacts [1]**	**DF Detecting using Eye Blinking [2]**	**DF detection by capsule networks [3]**
Key Identification	Face Wrapping artifacts	Involuntary eye blinking movements	Inbuild capsule networks
Technology used	Neural Networks	Neural Networks	Neural Networks
Accuracy	This model is tested for moderate to low accuracy	This model is tested for medium accuracy	This has the highest accuracy among all
Consumption of resources	High consumption of computing resources	Moderate to High consumption of computing resources	Highest consumption of resources accounting a proprietary capsules.

The Biological Signals Approach for Detecting Synthetic Portrait Videos extracts biological signals while performing authentic and fraudulent portrait video pairings, such biological signals can be gained from facial areas. Train a probabilistic SVM and a CNN using temporal consistency, capture signal characteristics in feature sets, convert to compute spatial coherence and PPG, and acquire the signal properties in advancement sets. Checks whether the video is real or not. [10]

Cele-DF is another project created to detect the deepfake video but it predicts based on the low resolution by improving the synthesized face to 256 × 256 pixels, check the colour mismatch, and by reducing the temporal flicking of the fake videos. [11]

In Effective and Fast Deepfake detection method based on Haarwavelet Transform, this project finds the deepfake by haar wavelet transform. It works by retrieving sharpness from blur pictures, edges of the images and also the synthesized surrounding area by using the haar function. They have used the UADFV dataset which basically have about 49 fake and 49 real videos. The accuracy proposed by these was 90.5%. It also works on the videos frames and each frame is inspected and face surroundings are extracted. [12]

III. PROPOSED SYSTEM

Many ways are present to make DeepFake videos but to detect them there are only a few possible ways. The technique used here to detect DF will secure the internet from the spread of DeepFake videos. The project provides a Django application that enables the users to upload the videos and justify whether it's real or fake. The project provides a web-based platform via the browser plugins made available for the DF detections.

The project can be incorporated in various applications such as WhatsApp and Facebook for identifying the DF videos before transferring them to a connected group or individual. The significant goal that can be achieved is in its performance and reliability in terms of usability, accuracy, reliability, and security.

The technique used here specializes in identifying various types of DeepFake like retrenchment DeepFake, replacement DeepFake.. Figure 4 represents the system architecture.

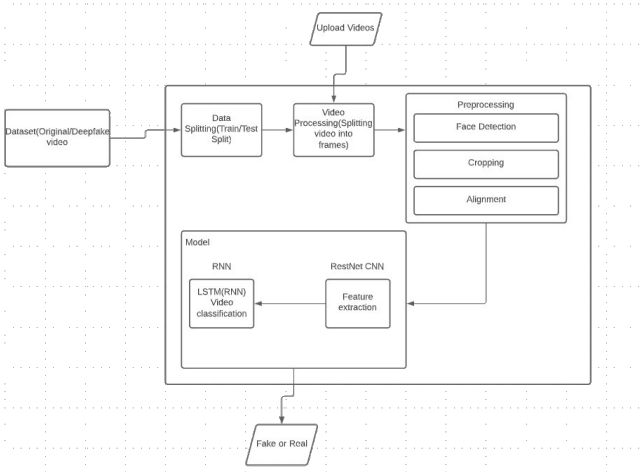

Fig. 4. System Architecture

A. Dataset Used:

The dataset used here is mixed with videos of equal numbers from various sources like Face Forensics++ followed by YouTube, various challenges datasets. By considering all these videos a new dataset is made which has fifty percent of the first video and fifty percent of tampered DF videos. Finally, the dataset is made by an appropriate division of thirty percent for test and seventy percent for the train.

Table 2 shows the total dataset created for deepfake till date.

Table 2. Basic details of the version dataset based on DeepFake [11]

Dataset	# Real		# DeepFake		Release Date
	Video	Frame	Video	Frame	
UADFV	49	17.3k	49	17.3k	2018.11
DF-TIMIT-LQ	320*	34.0k	320	34.0k	2018.12
DF-TIMIT-HQ			320	34.0k	
FF-DF	1,000	509.9k	1,000	509.9k	2019.01
DFD	363	315.4k	3,068	2,242.7k	2019.09
DFDC	1,131	488.4k	4,113	1,783.3k	2019.10
Celeb-DF	590	225.4k	**5,639**	2,116.8k	2019.11

B. Preprocessing Part

The pre-processing done on the dataset includes various steps like the splitting video into number of frames, detecting the face in frame and then cropping only the face part from it. In order to maintain consistency within the total count of frames, mean is been calculated of the dataset video, and also a new dataset is created which be having the cropped face. This dataset will have frames which will be equal to the mean calculated earlier.

The pre-processing part ignores some frames like which don't have face in it. As we know that if we process a 10 second video at frame count of 30 per seconds, i.e., about 300 frames, then a huge ton of computational power is been required. So for our hardware requirement match we have trained the model with only the first 100 frames.

C. Model Used

There are many parameters to consider a single layer of LSTM with resnext50 in the model creation. The Data Loader loads the preprocessed face-cropped videos and divides the videos into a test set and train set. Frames gained after preprocessing of the videos are provided to model for testing and training. Table 3 show error rate.

Table 3. Error rate of various models [13]

model	top-1 err.	top-5 err.
VGG-16 [41]	28.07	9.33
GoogLeNet [44]	-	9.15
PReLU-net [13]	24.27	7.38
plain-34	28.54	10.02
ResNet-34 A	25.03	7.76
ResNet-34 B	24.52	7.46
ResNet-34 C	24.19	7.40
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

D. Feature Extraction using ResNext CNN

The use of ResNext CNN in our project is to extract the features and accurately getting the frame level features also. Our network is finely tuned by addition of extra layers and then choosing the best rate and precisely converge the model gradient descent. Once the last pooling layer is completed, feature vector of size 2048-dimensional feature is created and it is used as the input of LSTM.

E. Sequence Processing using LSTM

Now, if we take for an example, 2 nodes of neural node and sequence of feature vectors of ResNext CNN of the frames as the input with the probability of sequence which are of the part of untampered or deepfake video. Then main moto of ours is to address the design of the model to continuously process the sequence in the proper sequential order. Figure 5 represents the LSTM for sequence processing.

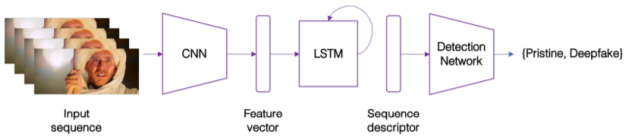

Fig 5. LSTM for Sequence Processing [9]

For achieving this, we are using the 2048 LSTM unit with dropout of 0.4 chance. To process the frames in sequential manner we make use of LSTM so that we do the temporal analysis of the video, by doing a comparison between the frames present at ‘t’ and ‘t-n’ second. Here ‘n’ represents the total number of frames present before t.

F. Model Accuracy based on frame count

Table 4. Model Accuracy

No of videos	No of frames	Accuracy
6000	10	84.21461
6000	20	87.79160
6000	40	89.34681
6000	60	90.59097
6000	80	91.49818

G. Prediction

To do the prediction of the video, an input video must be passed to trained model. The trained model mainly takes a pre-processed video as input and hence it must be in the same format. Further this video gets split into frames and then face cropping part is done.

Instead of using the local storage and occupying the memory by saving the video, there is a better way of directly passing the cropped frames to trained model. The output given by the model will be the video confidence related to the deepfake part along with the details of video like whether is real or fake one. Figure 6 represent Training flow.

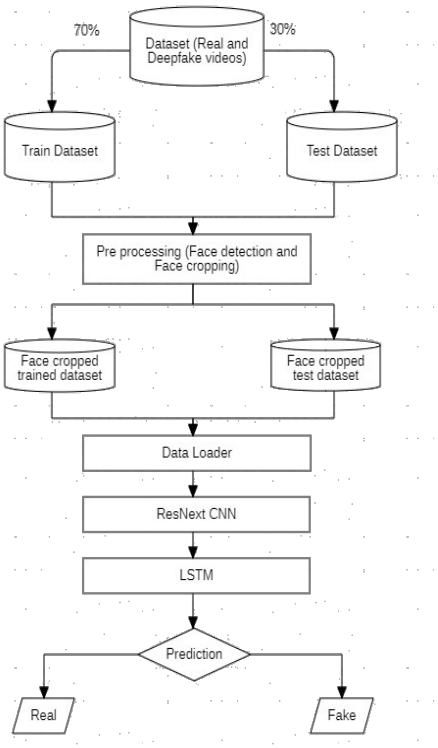

Fig. 6. Training Flow

IV. RESULT

A GUI will be provided for the user to upload the video and also, they can set the size of frame like 20,40,60,80. Uploading the corrupted video, long length videos, images other will give an error.

The final output will be the confidence level of model along with video detection as fake or real. Figures shown below depicts such occurrence. Outputs are shown in Figure 7, Figure 8, Figure 9. Prediction flow diagram is shown in Figure 10. Figure 11 shows the confidence of the video.

Fig. 7. GUI for video uploading

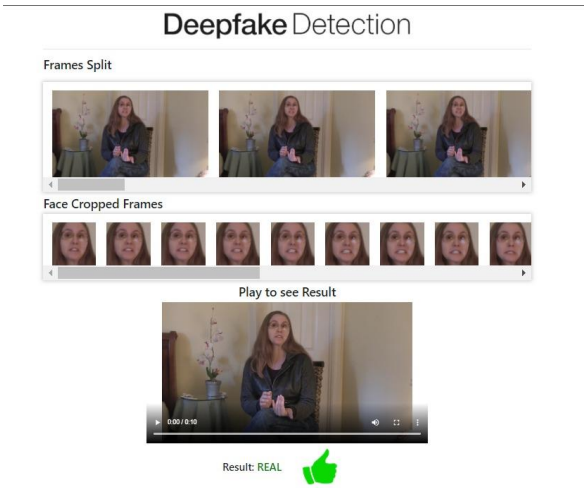

Fig. 8. Detection of Real Video

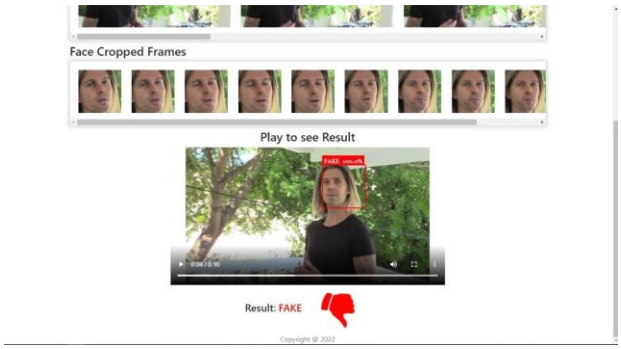

Fig. 9. Detection of Fake video along with confidence

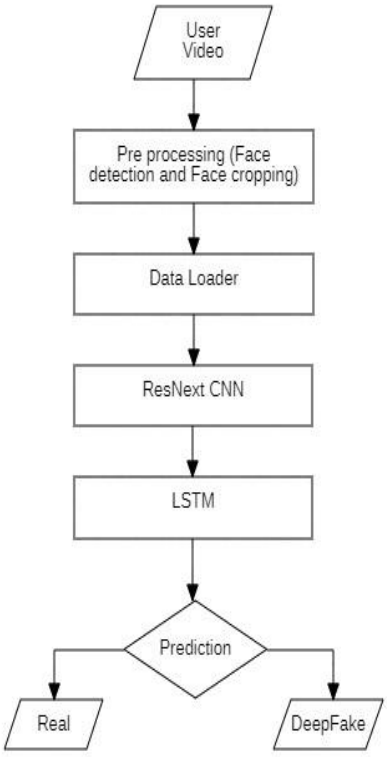

Fig. 10. Prediction Flow

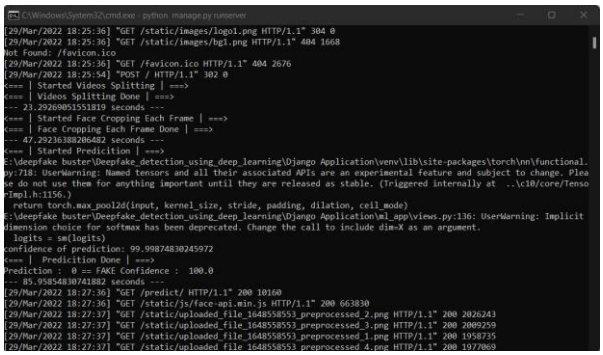

Fig. 11. Confidence of the video

V. CONCLUSION

We have come up with neural network-based approach for classifying the videos into the real or deepfake and showing the proposed model confidence. The proposed method is been created by keeping in mind the different ways of creating deepfakes using GANs and autoencoders. Our method make use of ResNext CNN for the frame level detection and RNN for the video classification and the LSTM also. The main goal of our project was to detect the accuracy of the video based on the alteration done and then classifying it as real or fake one based on certain parameters mentions in the paper. A high accuracy will be provided once the real time data will be used.

VI. LIMITATIONS

Our method has only considered the video part and not the audio. So, for that reason our method will not work for the audio deepfake. Hence, in future we can try to detect the audio deepfake done in the videos.

REFERENCES

[1] Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv conference May, 2019

[2] Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in IEEE conference 2018

[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”,IEEE conference, 2018.

[4] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in CVPR. IEEE, 2019.

[5] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv conference, May 2020

[6] MikaWesterlund, The Emergence of Deepfake Technology:, Technology Innovation Management, Version 9, November 2019.

[7] Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-ThecSaeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyen, arXiv conference, Febuary 2022.

[8] Rushikesh Potdar, Ajay Gidd, Shreya Kulkarni, Rohit Chavan, Prof. Nikam, International Research Journal of Modernization in Engineering Technology and Science, Volume:03/Issue:07/July-2021.

[9] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS conference, 2018.

[10] Umur Aybars Ciftci, ¨Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. X, No. X, July 2020

[11] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu, arXiv conference, arXiv:1909.12962v4, March 2020.

[12] Karthik P C, Sanjana S, M P Adithya Vijayan, Thushara P, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 10 Issue May-2021

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, arXiv conference, arXiv:1512.03385v1, Dec 2018.

[14] Luisa Verdoliva, Media Forensics and DeepFakes: an overview , arXiv:2001.06564v1, IEEE, January 2020.

[15] R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch, "Transferable deep-CNN features for detecting digital and print-scanned morphed face images," in CVPRW. IEEE, 2019.

[16] Tiago de Freitas Pereira, Andr´e Anjos, Jos´e Mario De Martino, and S´ebastien Marcel, "Can face anti spoofing countermeasures work in a real world scenario?,"in ICB. IEEE, 2020.

[17] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in WIFS. IEEE, 2018

[18] Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-ThecSaeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyen, arXiv conference, Febuary 2022.

[19] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “Faceforensics++: Learning to detect manipulated facial images,”in The IEEE International Conference on Computer Vision (ICCV), October 2019.October 2019.

[20] Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo,Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2):357–370, 2018.