DeepFake Detection Using Artificial Neural Network

Bharath Radhakrishna

Student Number 210667521

MSc. Artificial Intelligence, QMUL

ec211168@qmul.ac.uk

Project Supervisor: Eranjan Padumadasa

e.u.padumadasa@qmul.ac.uk

Abstract- Free deep learning-based software solutions have made it easier recently to create reliable face video interactions with little evidence of tampering, in the so-called "Deep Fake" (DF) videos. Digital video manipulations have been shown for, the availability of false content and the settings where it may be produced have greatly increased as a result of recent developments in the field of deep-learning. This has been accomplished by the skilful use of visual effects. These purportedly produced by AI-Media. The task of developing the DF using artificial intelligence methods is simple. However, it is difficult to locate these DF. because it takes a lot of work to train the algorithm to find the DF. We've made progress in locating the DF using convolutional neural networks and recurrent neural networks. Convolutional neural networks are used in this method to retrieve frame-level properties. A recurrent neural network that can detect whether a video is real-time or not is trained using these characteristics. Expected outcome when compared against a sizable collection of phoney videos obtained from a standard-data source. We use a straightforward design to show how our technology may be effective in this field.

Index Terms-Deep-Fake (DF),Long-Short-Term-Memory(LSTM),Convolutional-Neural-Network(CNN),Recurrent-Neural-Network(RNN)

I. Introduction

The making and distribution of digital movies is now easier than ever because to the improving smartphone cameras, the widespread availability of fast internet connections, and the ever-expanding reach of social media.

Deep learning has become so strong because to the increase in processing power that it was previously regarded to be unthinkable. This has brought forth new difficulties, as with any transformational technology. Deep generative adversarial models that have the ability to modify audio and video samples generate so-called "Deep Fake." It has been quite usual to spread the DF through social media platforms, which encourages spreading incorrect information and spamming. Those types of DF are terrible and lead to manipulative and threatening behaviour. These kinds of DF are horrible because they threaten and deceive the general public. DF detection is crucial for dealing with this problem. In order to efficiently identify between authentic videos and AI-generated phoney movies, we offer a novel deep learning-based technique (DF Videos). If the DF is to be located and its online spread stopped, methods that can identify fakes must be developed.

Understanding how Generative Adversarial Network (GAN) generates the DF is crucial for its identification. GAN accepts a video and a picture as input for a certain person (the "target") and produces a second video with a different person's face placed in place of the target's (the "source"). Deep adversarial neural networks, which automatically transfer the target's features and facial expressions as viewed from the source by training them on target movies and face photos, are the basis of DF. With the proper postprocessing, the created movies may achieve an unprecedented sense of authenticity. GANs is replaced in input picture in each frame after dividing the movie into frames. The video is further rebuilt. Usually, auto-encoders is utilised to execute the task. We present a brand-new deeplearning depend on approach that successfully separates DF videos from actual ones.

Proposed methodology is built on the same method used to produce Deep Fake using GAN.The procedure is based on a feature in DF movies that, owing to resource and time constraints, can only synthesise face pictures of a set pattern and that must go through an affinal-warping process in order to suit the source facial characteristics. Because the resolution differences between the surrounding environment and the distorted face-area, the final deepfake video has some discernible artefacts as a result of the warping. By breaking the video down into frames, using a ResNext Convolutional-Neural-Network to take the characteristics out, and detecting spatial abnormalities with both sequences presented from GAN throughout the regeneration of the DeeFake by a Recurrent Neural Network along

Long Short Term Memory, our method can identify these artefacts. We can train the ResNext CNN model more quickly by explicitly modelling the resolution disparity in affine face wrappings.

II. Related Work

Democracy, justice, and public trust are seriously threatened by the deep fake video industry's tremendous expansion and criminal usage. The need for bogus video analysis, identification, and action has risen as a result. The following is a list of some works that are connected to deep fake detection:

Exposing DF Videos by Detecting Face Warping Artifacts [1] employed a custom Convolutional-Neural-Network model to compare the produced face areas with their surrounding regions in order to detect artefacts. In this experiment, there were two different kinds of facial artefacts. Their strategy is based on the realisation that the deep-fake technique as it is can only create images with a limited resolution, that then must be further modified to correspond with the faces to be replaced in the original video. Temporal analysis of the frames was not considered in their methodology.

Exposing-AI-Created-Fake-Videos-by- Detecting-Eye-Blinking [2] proposes a novel technique for determining if a video is deep faked or authentic by using the eye blinking as a key feature. The clipped eye blinking frames underwent temporal analysis using the Long-term Recurrent Convolution Network (LRCNIt discusses a cutting-edge method for exposing fake face videos created by deep learning algorithms. The method is based on identifying eye flashing in videos, a physiological sign that is poorly represented in the fake, synthetic movies. The method's performance in detecting movies created by Deep Neural Network-based software DF is assessed over benchmarks of eye-blinking detection datasets and shows promise. Only the absence of blinking is used by their approach as a detection hint. However, additional factors like tooth enchantment, facial wrinkles, etc. must be considered for the detection of the deep fake. The suggested technique considers all of these factors. Since today's deepfake creation algorithms are so advanced, the absence of eye blinking cannot be the only indicator of a deepfake. For the detection of profound fakes, other factors like tooth enchantment, facial wrinkles, incorrect brow positioning, etc.

Using the capsule-network to detect-forged videos and images [3] it employs the technique that makes use of capsule-network to find faked, altered a range of images and films in various settings, such is the identification of computer-generated videos and replay attacks. They made a terrible decision by using random noise in the approach's training phase.. Even so, the model showed promise in their dataset, but because of training-stage noise, it might not work well on real-time data. It is suggested that our approach be trained on realtime, noiseless datasets. Recurrent Neural Network (RNN) for deepfake detection employed the strategy of employing RNN for sequential frame processing in addition to the pre-trained ImageNet model. They used the HOHO dataset, which only included 600 videos. Their dataset only has a tiny number of the same kind of films, which might make it difficult to perform effectively on real-time data. We'll use a lot of real-time data to train our model.

Detection-of-Synthetic-Portrait-Videos-using-Biological-Signals [4] Approaches using real and fake portrait video pairs extract biological signals from face areas. Apply changes to feature sets and PPG maps to capture signal properties, compute spatial coherence and temporal consistency, train CNN and probabilistic-SVM. The odds of the video's overall authenticity are then used to determine whether it is real or not. Independent of the generator, content, resolution, and video quality, Fraudulent Catcher accurately identifies fake material. It is not easy to create a variational loss-function that adheres to the recommended signal processing techniques since the discovery of biological signals is lost in the absence of a discriminator.

Using traditional neural networks, a unique method of comparing created face regions and their surrounding regions has been proposed. By Y. Li and S. Lyu [6]. The approach was founded on see whether the DF algorithm can produce photos with low resourcesI. Demir, L. Yin, and U. A. Ciftci's [7] focus was on feature extraction before computing temporal consistency and coherence. From the false and actual video pairs, the technique collected biological signals from face areas. A CNN and an SVN have been taught to determine the likelihood that something is real. Deepfakes may be automatically detected using the identification pipeline developed by D. Guera and J. Delp [8]. They have a recommended two-step analysis.. In the first-stage, CNN is used for extractaction of characteristics at the frame-level. In the second step, RNN will be used to collect irregular frames brought on by the face-swapping procedure. The dataset, which consists of 600 movies gathered from different web sources, was examined. Their model attained a 94 percent accuracy rate.

A novel method of uncovering deep fakes has been developed by Y. Li, MC. Chang, and S. Lyu [9] based on eye blinking patterns produced by neural networks. Since eye blinking is a natural indication and cannot be effectively communicated in minimal medium, in this study, we examined how it is represented in the video. The movies were initially pre-processed to identify the facial region in each frame. The temporal incongruity is then discovered

using a Long Term Recurrent Convolution Network (LRCN). Encoder - decoder architecture is the fundamental design for creating deep fakes. The encoder acquires the features of the target and source faces, and the decoder's job is to obtain the encoding features of the target face and then produce fake video.[10]. High level processing improves the video's quality and gets rid of the residues, but there are still a few traces that aren't apparent to the human eye. The main characteristics of our detection model are these lingering residues.[11]. The InceptionResnetV2 algorithm is used in the suggested model to extract features. These characteristics are then used to train a recurrent neural network, which determines whether or not the video has been altered. The video is divided into small frames and these frames are used as input to the detection model since only a little section of the video is altered, making the deepfakes shorter in duration. [14].

III. Methodology

Although there are several tools accessible for DF creation, there are very few ones available for DF detection. Our method of DF detection will significantly contribute to preventing the spread of the DF over the internet.

We'll provide a website where people may submit videos and identify them as real or fake. This project may be built up to include developing a web-based platform and a browser plugin for automatic DF detections. Even well-known applications like Facebook and WhatsApp may include this project into their code for straightforward DF pre-detection before forwarding to another user. One of the main objectives is to gauge how well it performs and is accepted in terms of security, usability, correctness, and dependability.

Our approach focuses on identifying all forms of DF, including interpersonal, replacement, and retrenchment DF. The suggested system's basic system architecture is shown in figure 1: -

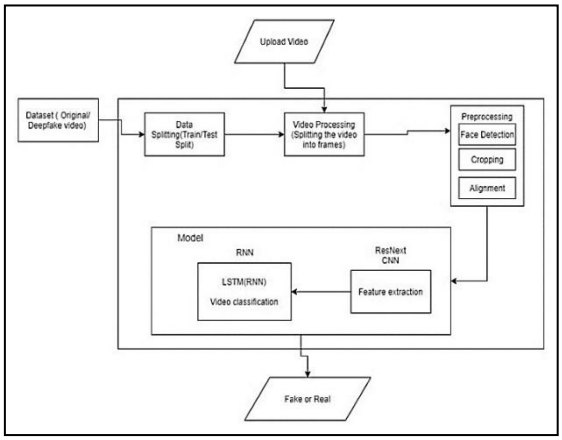

Figure1:- System-Architecture

i. DATABANKS

A mixed dataset that we are using consists of an equal number of videos from FaceForensics++[14], YouTube, and the Deepfake-Detection-Challenge-Dataset[13].

50 percent of original-video and 50 percent of updated DF movies are included in our recently generated dataset. From the dataset, two sets are produced: a test set and a train set.

ii. Initialisation

The movie is divided into frames as part of the dataset preparation procedure.

The frame is then trimmed to include the identified face. A new processed face cropped datasets is produced utilising the frames that make up the mean of the dataset video in order to ensure consistency in the number of frames. Pre-processing does not take the frames with no faces into consideration.

Processing the 300 frames in a 10 second movie at 30 frame per second will require a lot of processing power. Therefore, we advise that the model be trained using just the first 100 frames for experimental reasons.

iii. Model

The model comprises of one LSTM layer followed by resnext50 32x4d. The pre-processed face-cropped movies are loaded by the data loader, who divides them into a train set and a test set. Additionally, the model receives small batches of the altered film frames for testing and training.

iv. For Feature Extraction, use ResNext CNN

We suggest using ResNext CNN divider for properly recognising the frame level features rather than constructing the classifier from scratch in order to extract the features. The network will then be fine-tuned by adding any additional necessary layers and choosing an appropriate learning rate to correctly converge the gradient descent of the model.

Following the last pooling layers, the 2048 dimensional feature vectors that are still present are used as the sequential LSTM input.

v. LSTM-Based Sequence Processing

Think about an input consisting of a series of ResNext CNN feature vectors for a two-node neural network. Determine how likely it is that the sequence comprises unaltered or deeply faked footage. The key problem that has to be solved is how to construct a model that meaningfully process a sequence recursively. To achieve this objective, we suggest the need of 2048-LSTM along a (0.4) danger in dropouts for this problem.

The frames are successively examined using LSTM in order to do a temporal analysis on the video and compare the frame at second t with the frame at second t-n. where n is the number of frames prior to frame t.

vi. Forecast

The trained algorithm is given a fresh video to forecast using. Additionally, a fresh video is pre-processed to incorporate the trained model's format. Instead of being locally saved, the video is divided into fragments, the faces are cropped, and the segments are then sent immediately to the trained-model for face recognition.

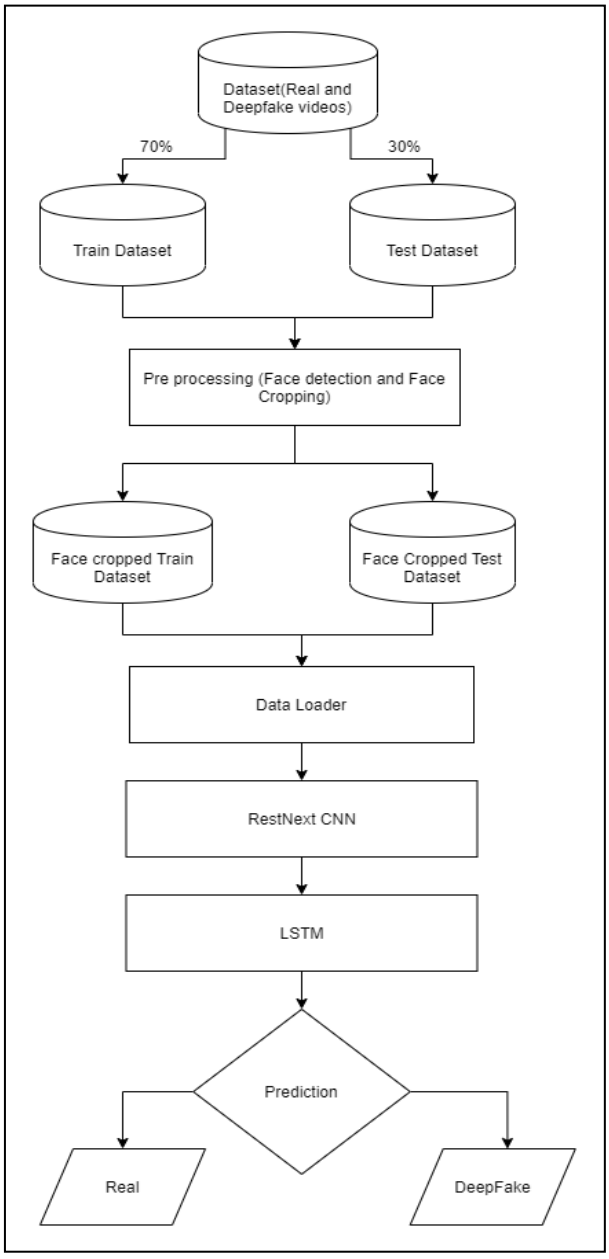

Figure 2:-TrainingFlow

IV. OUTCOME

Model's output will include model's confidence level and a determination of whether movie is authentic or a deepfake. Figure 3 presents one instance.

Figure 3:- Expected-Outcome

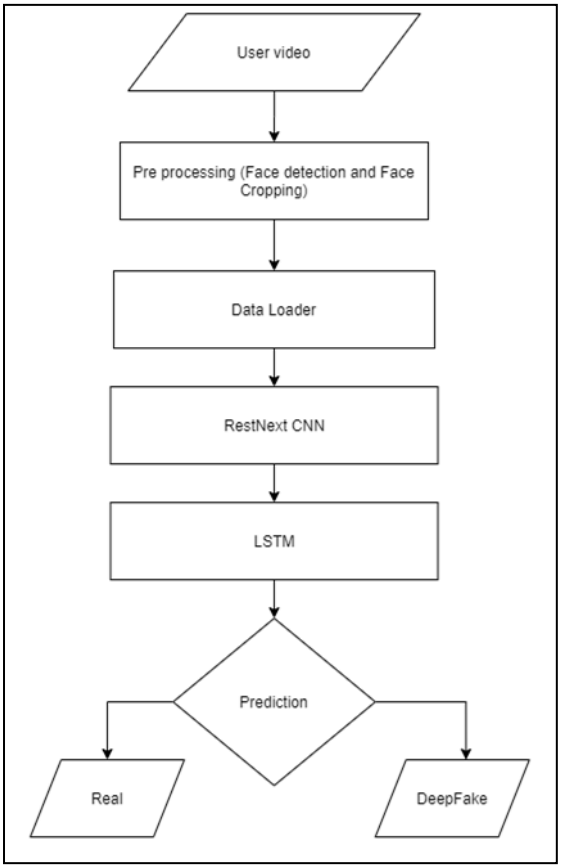

Figure 4:-Prediction-Flow

Model Name	Dataset	No. of videos	Sequence length	Accuracy
model_90_acc_20_frames_FF_data	FaceForensic++	2000	20	90.95477
model_95_acc_40_frames_FF_data	FaceForensic++	2000	40	95.22613
model_97_acc_60_frames_FF_data	FaceForensic++	2000	60	97.48743
model_97_acc_80_frames_FF_data	FaceForensic++	2000	80	97.73366
model_97_acc_100_frames_FF_data	FaceForensic++	2000	100	97.76180
model_93_acc_100_frames_celeb_FF_data	Celeb-DF + FaceForen-sic++	3000	100	93.97781
model_87_acc_20_frames_final_data	Our Dataset	6000	20	87.79160
model_84_acc_10_frames_final_data	Our Dataset	6000	10	84.21461
model_89_acc_40_frames_final_data	Our Dataset	6000	40	89.34681

Table 1: Trained-Model Results

V. CONCLUSION

The classification of the video as deep fake or real using a neural network-based method was provided, along with the suggested model's level of confidence. The DeepFakes made with the aid of GANs are the model that the suggested strategy is based on. The DeepFakes generated using GAN with the assistance of Autoencoders served as motivation for the proposed tactic. For scene detection, our method leverages ResNext CNN, and for video-classification, RNN and LSTM. Based on the criteria outlined in the research, the suggested approach can determine if a film is DeepFake or real. We believe it will provide incredibly precise real-time data.

VI. DRAWBACKS

We did not account for the audio in our technique. Due to this, the audio deep fake cannot be detected by our approach. But in the future, we suggest achieving the identification of audio deep fakes.

VII. REFERNCES

1) Li, Y. and Lyu, S. (2019). Exposing DeepFake Videos By Detecting Face Warping Artifacts. *CVPR Workshops*. [online] Available at: https://www.semanticscholar.org/paper/Exposing-DeepFake-Videos-By-Detecting-Face-Warping-Li-Lyu/2d066beb34469559e0fc5e5ab4d68dc736cfd46f.

2) Li, Y., Chang, M.-C. and Siwei Lyu (2018). *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking*. [online] undefined. Available at: https://www.semanticscholar.org/paper/In-Ictu-Oculi%3A-Exposing-AI-Generated-Fake-Face-by-Li-Chang/47e8acbabda89b0ece513bd90d0b669d9385fcc3 [Accessed 17 Oct. 2019].

3) Nguyen, H.H., Yamagishi, J. and Echizen, I. (2019). Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

doi:10.1109/icassp.2019.8682602.

4) Ciftci, U.A. and Demir, I. (2020). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [online] pp.1–1. doi:10.1109/TPAMI.2020.3009287.

5) Kim, H., Theobalt, C., Carrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C. and Zollhöfer, M. (2018). Deep video portraits. *ACM Transactions on Graphics*, [online] 37(4), pp.1–14. doi:10.1145/3197517.3201283.

6) Ciftci, U.A. and Demir, I. (2020). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [online] pp.1–1. doi:10.1109/TPAMI.2020.3009287.

7) Guera, D. and Delp, E.J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018*

15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). doi:10.1109/avss.2018.8639163.

8) Li, Y., Chang, M.-C. and Lyu, S. (2018). *In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking*. [online] IEEE Xplore. doi:10.1109/WIFS.2018.8630787.

9) Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). *Generative Adversarial Nets*. [online] Neural Information Processing Systems. Available at: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

10) He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778. doi:10.1109/cvpr.2016.90.

11) Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C. and Mallya, A. (2021). Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications. *Proceedings of the IEEE*, pp.1–24. doi:10.1109/jproc.2021.3049196.

12) Bouarara, H.A. (2021). Recurrent Neural Network (RNN) to Analyse Mental Behaviour in Social Media. *International Journal of Software Science and Computational Intelligence*, 13(3), pp.1–11. doi:10.4018/ijssci.2021070101.

13) Raghavendra, R., Raja, K.B., Venkatesh, S. and Busch, C. (2017). Transferable Deep-CNN Features for Detecting Digital and Print-Scanned Morphed Face Images. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi:10.1109/cvprw.2017.228.

14) Singh, R.K., Sarda, P.V., Aggarwal, S. and Vishwakarma, D.K. (2021). Demystifying deepfakes using deep learning. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. doi:10.1109/iccmc51019.2021.9418477.

15) Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A. (2017). *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. [online] www.aaai.org. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14806.

16) Guera, D. and Delp, E.J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. doi:10.1109/avss.2018.8639163.

17) PyTorch Forums. (2017). *Confused about the image preprocessing in classification*. [online] Available at: https://discuss.pytorch.org/t/confused-about-the-imagepreprocessing-in-classification/3965 [Accessed 15 Aug. 2022].

18) Qian, Y., Chen, K., Nikkanen, J., Kämäräinen, J.-K. and Matas, J. (n.d.). *Recurrent Color Constancy*. [online] Available at: https://openaccess.thecvf.com/content_ICCV_2017/papers/Qian_Recurrent_Color_Constancy_ICCV_2017_paper.pdf [Accessed 15 Aug. 2022].

19) Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A.A. (2017). *Image-To-Image Translation With Conditional Adversarial Networks*. [online] openaccess.thecvf.com. Available at: https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html.

20) Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A.A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [online] doi:10.1109/cvpr.2017.632.

21) Rahmouni, N., Nozick, V., Yamagishi, J. and Echizen, I. (2017). *Distinguishing computer graphics from natural images using convolution neural networks*. [online] IEEE Xplore. doi:10.1109/WIFS.2017.8267647.

22) Song, F., Tan, X., Liu, X. and Chen, S. (2014). Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. *Pattern Recognition*, 47(9), pp.2825–2838. doi:10.1016/j.patcog.2014.03.024.

23) King, D.E. (2009). Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* [online] doi:10.5555/1577069.1755843.