Development of AI/ML based solution for detection of face-swap based deep fake videos

Naga Nikhitha Peddisetty  
*SoCSE & IS*  
*Presidency University  
Bangalore, India*  
[*naganikithapeddisetty@gmail.com*](mailto:naganikithapeddisetty@gmail.com)

Sarthak Mishra  
*SoCSE & IS*  
 *Presidency University  
Bangalore,India*  
[*mishrasarthak052002@gmail.com*](mailto:mishrasarthak052002@gmail.com)

*Dr. Mohammedi Akheela Khanum*  
*SoCSE & IS*  
*Presidency University  
Bangalore, India*  
[*akheela.khanum@presidencyuniversity.in*](mailto:akheela.khanum@presidencyuniversity.in)

Darshan S  
*SoCSE & IS*  
 *Presidency University  
Bangalore,India*  
 [*darshan110604@gmail.com*](file:///C:\Users\Naga%20Nikitha\OneDrive\Desktop\a2sl\darshan110604@gmail.com)

*Abstract*—within recent months, free tools based on deep learning have made it easier to produce genuine face exchanges in videos that bear minimal traces of manipulation, in so-called "DeepFake" (DF) videos. Digital video manipulations have been shown for decades through the proper utilization of visual effects, but recent breakthroughs in deep learning have caused a sharp rise in the authenticity of synthetic content and the ease with which they can be developed. Such so-called AI-synthesized media (commonly known as DF). Developing the DF through artificially intelligent tools are easy task. But, when it comes to the detection of the DF, it is a big challenge. Because the training of the algorithm to identify the DF is not easy. We have moved one step ahead in identifying the DF with Convolutional Neural Network and Recurrent Neural Network. The system employs a convolutional Neural network (CNN) to extract frame-level features. These features are utilized to train a recurrent neural network (RNN) which learns to classify whether a video has been manipulated or not and is capable of identifying the temporal inconsistencies between frames brought by the DF creation tools. The expected outcome against a huge set of fabricated videos gathered from the normal data set. We demonstrate how our system can be competitive outcome in this task leads to utilizing a simple architecture.

Keywords—Deepfake Video Detection, convolutional Neural network (CNN), recurrent neural network (RNN).

# Introduction

In The progress in smartphone camera technology, coupled with the widespread availability of reliable internet access globally, has significantly expanded the influence of social media and media sharing platforms. This evolution has simplified the creation and distribution of digital videos compared to previous times. Advancements in computational power have made deep learning much more powerful, to a degree that would have seemed unrealistic just a few years ago. However, like any groundbreaking technology, it has introduced new challenges. Notably, "DeepFake" content generated by deep generative adversarial models can manipulate video and audio segments. The proliferation of such deepfakes on social media has become widespread, leading to the dissemination of misleading information and spam across these platforms. These types of deep fakes are problematic, posing threats and potentially misleading everyday users.

To address this issue, it is crucial to develop effective deepfake detection methods. We propose a novel approach based on deep learning that can accurately distinguish AI-generated fake videos from genuine ones. The ability to identify these fakes is essential in preventing their spread online.

Understanding how Generative Adversarial Networks (GANs) create deepfakes is vital for detection. GANs take a video and a picture of a target person and generate a new video that replaces the target's face with that of another individual. The essence of deepfakes lies in deep adversarial neural networks trained with facial images and target videos, allowing for the automatic mapping of facial features from the source to the target. With proper post processing, the result can appear highly realistic. The GAN operates by breaking the video into frames, substituting the input image in each frame, and then reconstructing the video, often using autoencoders.

Our deep learning-based approach effectively distinguishes deep fake videos from real ones by leveraging the same principles employed by GANs in their generation. This method capitalizes on specific characteristics of deep fake videos; due to limitations in computational resources and production time, the algorithm can only synthesize face images of fixed sizes, necessitating affine warping to match the source's facial structure. This warping leads to noticeable artifacts in the final deepfake video, especially due to resolution discrepancies between the altered facial region and its context.

# LITERATURE SURVEY

The rapid increase in the use of deep fake videos, particularly for malicious purposes, poses a significant danger to democracy, justice, and public confidence. Consequently, there is a growing need for analysis, detection, and intervention related to fake videos. Below are some terms associated with deep fake detection:

Detecting Face Warping Artifacts in Deep Fake Videos [1] describes an approach that identifies artifacts by comparing the areas of generated faces to their adjacent regions using a specialized Convolutional Neural Network model. This study focuses on two types of face artifacts. The method is grounded in the observation that current deep fake algorithms can only produce images with limited resolutions, which must then be transformed further to align the faces with those in the original video.

Identifying AI-Generated Fake Videos through Eye Blinking Detection [2] presents a novel technique for revealing fake face videos created with deep neural network models. This approach is based on detecting eye blinking in the videos, as this physiological indicator tends to be absent or poorly represented in the artificially generated content. The method has been evaluated using benchmarks from eye-blinking detection datasets and demonstrates effectiveness.

Their approach relies solely on the absence of blinking as an indicator for detection. Nevertheless, additional factors need to be taken into account for identifying deep fakes, such as dental alterations, facial wrinkles, and so forth. Our method is designed to incorporate all these variables.

The use of capsule networks to identify manipulated images and videos employs a technique that utilizes a capsule network for detecting forged or altered images and videos across various scenarios, including replay attack identification and computer-generated video detection.

In their strategy, random noise was introduced during the training phase, which is not an ideal choice. Although the model yielded beneficial results on their dataset, it may encounter difficulties with real-time data due to the training noise. Our method aims to be trained on datasets that are both noiseless and representative of real-time conditions.

The method for detecting synthetic portrait videos by analyzing biological signals extracts biological indicators from facial areas in both genuine and counterfeit portrait video pairs. It performs transformations to evaluate spatial coherence and temporal consistency, captures signal characteristics in feature sets and PPG maps, and trains a probabilistic SVM alongside a CNN. Ultimately, it aggregates the authenticity probabilities to determine whether the video is genuine or fake.

Fake Catcher identifies fake content with high precision, regardless of the generator, type of content, resolution, or quality of the video. However, the absence of a discriminator has led to losses in their findings concerning the preservation of biological signals, making the formulation of a differentiable loss function that adheres to the proposed signal processing steps a complex task.

# PROPOSED SYSTEM

Numerous tools exist for constructing the DF, yet there are very few options available for its detection. Our methodology for identifying the DF will significantly aid in preventing its spread across the internet. We will offer a web-based platform that allows users to upload video content and categorize it as either fake or genuine. This initiative has the potential to evolve from a web-based platform into a browser extension for automatic DF detection. Major applications like WhatsApp and Facebook could incorporate this project into their platforms to facilitate preliminary detection of DF before it is shared with other users.

A key goal is to assess its effectiveness and user acceptance in terms of security, ease of use, precision and dependability. Our approach aims to detect all forms of DF, including replacement DF, retrenchment DF, and interpersonal DF. Figure 1 illustrates the basic system architecture of the proposed solution.

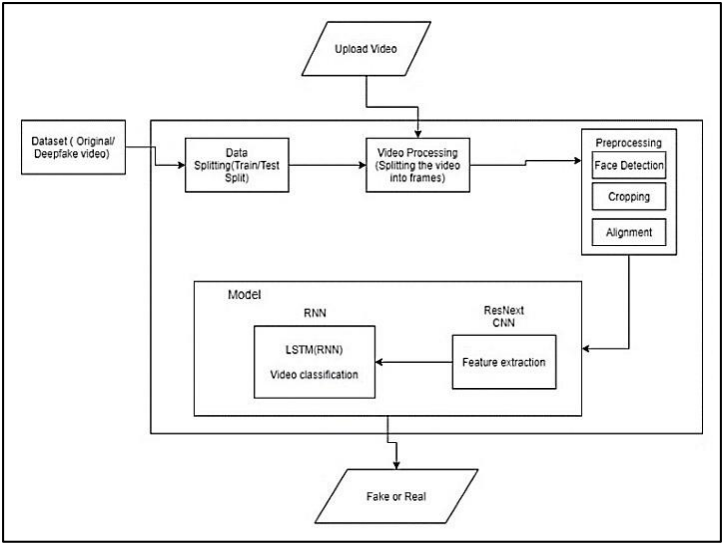


Fig.1. System Architecture

## Dataset

We are utilizing a diverse dataset that includes an equal number of videos sourced from various platforms such as YouTube, FaceForensics++, and the deepfake detection challenge dataset. Our newly compiled dataset comprises 50% original videos and 50% manipulated deepfake footage. This dataset is divided into a training set (70%) and a testing set (30%).

## Prepocessing

The preprocessing of the dataset involves dividing the video into individual frames. This is followed by detecting and cropping faces from these frames. To ensure uniformity in the number of frames, we calculate the average number of frames in the dataset videos and create a new dataset that contains the same number of cropped face frames. Any frames without detected faces are disregarded during this process. Processing a 10-second video at 30 frames per second, resulting in a total of 300 frames, requires significant computational resources. Therefore, for experimental purposes, we will only use the first 100 frames to train the model.

## Model

Our model is based on ResNext50\_32x4d, followed by a single LSTM layer. The Data Loader retrieves the preprocessed, face-cropped videos and divides them into training and testing sets. Subsequently, the frames from these processed videos are fed into the model in mini-batches for training and evaluation.

## ResNext CNN for Feature Extraction

Rather than developing a new classifier, we propose utilizing the ResNext CNN classifier to extract features and accurately identify frame-level characteristics. We will fine-tune the network by adding necessary layers and selecting an appropriate learning rate to ensure effective convergence of the model’s gradient descent. The resulting 2048- dimensional feature vectors from the final pooling layers will serve as the input for the sequential LSTM.

## LSTM for Sequence Processing

We will treat a sequence of ResNext CNN feature vectors from the input frames as input to a 2-node neural network that assesses the likelihood of the sequence belonging to a deepfake video or an authentic one. A significant challenge to address is the design of a model capable of processing this sequence meaningfully. To tackle this problem, we propose using an LSTM unit with 2048 units and a dropout rate of 0.4, which meets our objectives. The LSTM will process the frames sequentially, enabling temporal analysis by comparing the frame at time ‘t’ with the frame from ‘t-n’ seconds, where n can represent any number of frames preceding.

## Predict

To predict, a new video is introduced to the trained model. This video undergoes preprocessing to match the format required by the model. It is segmented into frames, followed by face cropping, and instead of saving the video locally, the cropped frames are directly submitted to the trained model for detection.

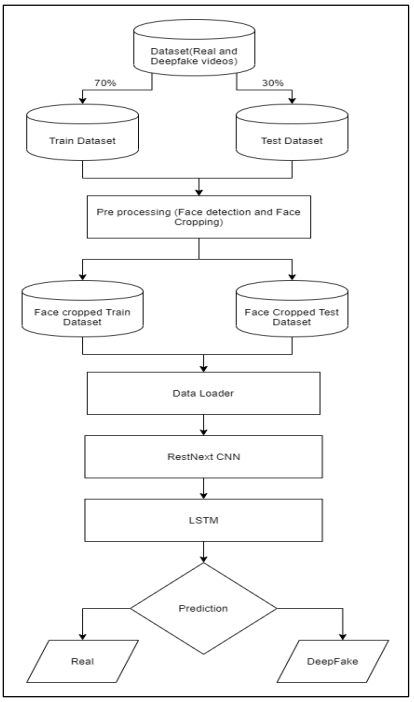


Fig.2.: Training Flow

# RESULT

The output of the model is going to be whether the video is deepfake or a real video along with the confidence of the model.

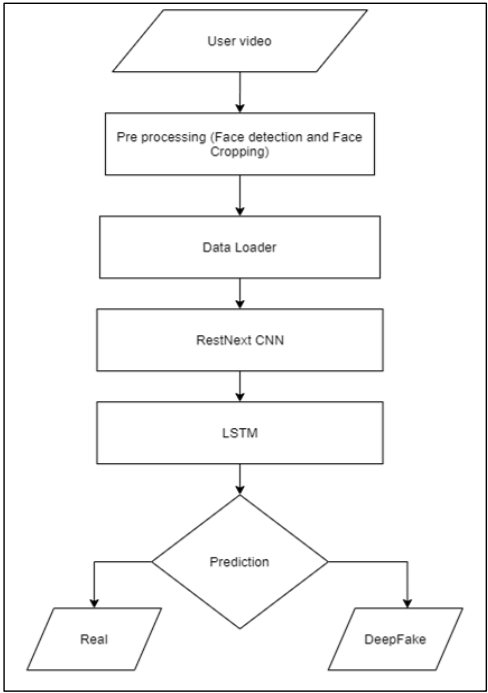


Fig.3: Prediction flow

# CONCLUSION

We introduced an approach that utilizes a neural network to determine whether a video is a deep fake or authentic, along with assessing the confidence of the model we proposed. This method draws inspiration from the creation of deep fakes through Generative Adversarial Networks (GANs), as discussed in "Deepfake Video Detection using Neural Networks" (IJSRD/Vol. 8/Issue 1/2020/221) Autoencoders. Our technique performs detection at the frame level by employing ResNext CNN and classifies the video using an RNN with LSTM. The method is designed to identify videos as either deep fakes or genuine based on the parameters outlined in the paper. We are confident that it will achieve very high accuracy with real-time data.

# LIMITATIONS

Our method has not considered the audio. That’s why our method will not be able to detect the audio deep fake. But we are proposing to achieve the detection of the audio deep fakes in the future.

REFERENCES

1. Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3.
2. Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.
3. Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.
4. Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv:1901.02212v2.
5. Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2.
6. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.
7. David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
9. An Overview of ResNet and its Variants : Vision and Pattern Recognition, pages 5967–5976, July 2017. Honolulu, HI
10. 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, 2017.

[11] 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, 2013.

[12] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, “Distinguishing computer graphics from natural images using convolution neural networks,” in WIFS. IEEE, 2017.

[13] F. Song, X. Tan, X. Liu, and S. Chen, “Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients,” Pattern Recognition, vol. 47, no. 9, pp. 2825–2838, 2014.

[14] D. E. King, “Dlib-ml: A machine learning toolkit,” JMLR, vol. 10, pp. 1755–1758, 2009.

[15] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in ICCV, 2019.

[16] Xinyuan Chen, Changxu Fan, and Xiaoming Liu, “SelfBlended Images for Detecting Image Forgeries,” in NeurIPS, 2021.

[17] Nhu-Van Nguyen, Fuming Fang, and Dacheng Tao, “Multi branch attention networks for deepfake detection,” in ICASSP, 2021.

[18] Tali Dekel, Michael Rubinstein, Ce Liu, and William T. Freeman, “On the effectiveness of visible watermarks,” in CVPR, 2017.

[19] Luigi Verdoliva, “Media forensics and deepfakes: An overview,” in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.

[20] Justus Thies, Michael Zollhöfer, and Matthias Nießner, “Deferred neural rendering: Image synthesis using neural textures,” in TOG (Proc. SIGGRAPH), 2019.

[21] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi, “Detection of GAN-generated fake images over social networks,” in IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018.

[22] Xin Yang, Yuezun Li, and Siwei Lyu, “Exposing deep fakes using inconsistent head poses,” in ICASSP, 2019.

[23] Yaojie Liu, Lingxi Xie, Xin Wang, and Alan Yuille, “Deep Learning Face Attributes in the Wild,” in ICCV, 2015.

[24] Ekraam Sabir, Jianying Zhou, Farhan Akhtar, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” in CVPR Workshops, 2019.

[25] Muhammad Asad, Adnan Qayyum, Imran Razzak, and Sung Wook Baik, “Deepfake video detection using efficientnet-v2 and attention mechanism,” in Sensors, 2021.

[26] H. Farid, “Image forgery detection,” IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 16–25, 2009.

[27] Zhiwei Yu, Jiangbin Zheng, Mengzhou Xia, and Chao Liu, “FakeLocator: Robust localization of GAN-based face manipulations via semantic segmentation networks with regional consistency,” in Pattern Recognition, vol. 122, 2022.