International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 6, Nov- Dec 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Development of AI/ML Based Solution for Detection of Face Swap Based Deepfake Videos

[1]Ms. Vineetha B, [2]Mandhara J, [3]Sayeed Ahmed CS, [4]Rayyan Shariff
1Assistant Professor, Department of Computer Sciences and Engineering, Presidency University, Bangalore, India
2,3,4Department of Computer Sciences and Engineering, Presidency University, Bangalore, India.

Abstract—

The increased complexity of face-swap deepFake videos makes them more difficult to differentiate with real videos and poses significant threats of misinformation and undermined reputation. This paper introduces an AI/ML-based solution built to detect such manipulations accurately and efficiently by spotting subtle artifacts. We propose a hybrid model that integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to capture both spatial and temporal patterns in video data. The CNN extracts spatial details from individual frames, such as visual artifacts, while the LSTM examines temporal inconsistencies between frames, like unnatural motions or expressions. This combined approach aims to deliver robust, adaptable model capable of achieving high accuracy in distinguishing authentic videos from deepfakes, helping preserve digital trust.

Keywords—Deepfake detection, CNN, LSTM, Face-swap, Misinformation, AI, ML, Video analysis.

I. INTRODUCTION

In the modern interconnected world, the problem with AI-created synthetic media, also known as as deepfakes, is very dangerous to the trustworthiness of visual information. Another type of fake content generated through deepfakes, known as face-swap deepfakes, wherein the face of one individual is subtly overlaid on the body of another is incredibly effective in terms of propagating fake content, identity theft, and causing significant reputational damage. Such manipulations are critical to the integrity of journalism, legal systems, and social media platforms and therefore need to be detected.

Conventional authentication approaches tend to rely on the human eye to detect hidden clues such as the presence of unnatural blinking or lights being incorrect; trying to do so is even more subjective, time consuming, and can be easily compromised. Initial forensic tools which depended on frequency or edge detection are now mostly useless in the face of more sophisticated machine generators like Generative Adversarial Networks (GANs), which generate extremely more lifelike content. As the application of AI tools to create deepfakes is now a common occurrence, the detection systems have to be implemented on a large scale and automated. This project will fulfill this need by creating a machine learning-powered system with the capability of providing real-time, reliable authenticity analysis.

II. LITERATURE REVIEW

Earlier studies have experimented with deepfake detection features with various methods, which a few are limited to achieve but due to the emergence of newer generative models, many are easily traced. The detection strategies presently in use can be divided into three basic groups, i.e. spatial, temporal and physiological analysis.

The area of interest of Spatial Analysis is determined by inconsistencies in individual video frames, e.g., inconsistency between head modes, unusual lighting, unnatural skin color, etc. However, where these methods are effective in terms of early deepfakes, they are much less effective when working with contemporary GANs that can produce more realistic and smooth images.

Temporal Analysis examines anomalies in time between one frame and the next - one issue around this is flickering or blinking of the swapped face irregularly over time. Although these methods of applying artifacts could help generate flaws, they usually focus on a particular vulnerability and can be evaded as newer deepfake generators create methods to remove these identifiable features.

Another emerging method is called Physiological Signal Detection and is aimed at detecting those inconspicuous changes in physiology (heart rates, etc.), which deepfakes may not sufficiently learn to imitate. These methods are however usually complicated and might not be applicable in real-time. The first deficiency in the existing literature is the lack of a single model that would incorporate both spatial and temporal analysis and construct a more powerful and generalized detection model.

III. PROPOSED METHODOLOGY

In the proposed methodology, we outline how we are going to develop an extensive system to identify face-swap deepfakes by analyzing the spatial and temporal characteristics of video information. It involves the selection of a dataset, feature extracts and pre-processing, model architecture and the performance measures used to evaluate it.

A. Dataset Used

Benchmark datasets widely accepted in the field including the ones at least will be used to train and test the models.

FaceForensics++ and the DeepFakeDetection Challenge dataset. These datasets provide large sizable groups of authentic and forged face-swap videos that are critical in developing a powerful classifier. The data will be divided into training and testing data to make sure that the

effectiveness of the model will be thoroughly evaluated on data that is not seen by it.

B. Representation and Semantic oddity, the Extraction of features.

Our system consists of a major component that is an optimized preprocessing pipeline, which is to prepare video data to be used in training and analysis in models.

1. Frame Extraction: It begins with the extraction of separate frames of each input video with the help of OpenCV library.

2. Face Detection and Cropping: Faces on each frame are detected and then there is face cropping of the frame using MTCNN (Multi-Task Cascaded Convolutional Networks) or using Haar Cascade as an alternative detector.

3. Image Resizing: The facial images which are cropped are resized to the same size of 160X160 pixels to ensure that all are the same size to be used as inputs to the model.

4. Sequence Preparation: The oracles have been finally prepared by organizing the processed frames into sequences and storing them through NumPy arrays (.npy files) to allow them to be loaded efficiently to train and make inferences with the model.

C. Extracting and Preprocessing of Features.

Our algorithm is based on a hybrid deep learning model based on a Convolutional Neural Network (CNN) and a Long Short-Term Memory network (LSTM).

- There are no publicly accessible component models, and their numbers vary with the video's resolution size.<
- CNN: CNN component does the spatial feature extraction, which analyzes single frames to detect subtle artifacts, compression artifacts or inconsistency in facial features, which is common in the deepfake video. These spatial features are capture with the help of conv2D layers.
- LMST: LSTM is an element, which locates the temporal connection through the analysis of the frame sequences. With TimeDistributed and LSTM layers it is able to detect irregularities present across frames e.g. unnatural motion or sudden changes that cannot be learned by the analysis of single frames.

This intermediate design allows the model to learn what patterns to identify in an image as well as the changing variable patterns over time which makes it more complete and precise detecting deep Fresh algorithms.

D. Performance Metrics

Various standard evaluation metrics will be used in the process of evaluating the model performance so as to achieve a dependable as well as an effective classification model.

- Accuracy: This is the percentage of videos correctly read as real or fake, and the goal accuracy test is 95 or more.
- Is a quality metric that assesses the extent to which a model can detect high quality positive (deepfake) examples, which limits the number of false positives.
- Detection all of the true positives and thus minimizing false negatives, thus the recollections indicate the ability of the model to distinguish between the true and false events.
- F1-Score: It is the harmonic average of the recall and the precision, and provides Model performance balanced measures.
- LSTM: LSTM is an element that takes into account temporal features through an analysis of frames. By employing TimeDistributed and LSTM layers, the model can point at the presence of frame-to-frame anomalies (e.g. unnatural movements or jump cuts) that are difficult to see when individual frames are analysed separately.

IV. Experimental Results

- With the suggested methodology, the system is expected to provide impressive gains to models that can only use the spatial or temporal analysis. The most important anticipated results are:
- In this case, the highest accuracy will be 95 or better on the held-out test set with the least false detections.
- Producing a 1520 percent accuracy improvement over the default CNN model that does not feature temporal analysis.
- Performing better and more adaptively to novel, newer techniques of deepfake generation than the previous ones, which is a widespread weakness of existing methodologies.
- Provision of sufficient processing speed to support real-time or near-real-time detection, so that the system is useful in analysis of live video.

V. System Workflow

In the entire system workflow, the following arrangement is made:

1) Data Acquisition: The system can accept video files in common formats (i.e. mp4).

Frame Extraction: The extraction of individual frames of the video is applied to OpenCV.

3) Face Detection and Cropping: MTCNN is used to detect faces in each frame and cropping as well as scaling face to 160 x 160 is used to generate uniform faces.

4) Sequence Preparation: The processed frames are formatted into fixed length sequences and saved in the form of.npy files.

5) Model Inference: These sequences are fed through theModel CNN+LSTM which examines both spatial as well as temporal patterns.

6) Evaluation and Output: The system produces a final model of either a real or a fake model with a confidence value indicating the level of certainty of the model.

VI. Development Technologies

This solution implementation is based on a contemporary, open-source technology stack which supports AI applications and machine learning.

- Programming Language Python 3.11 will be used as the main language of development, as it has a powerful ecosystem and extensive libraries based on AI, ML, and image processing applications.
- Deep Learning Frameworks: In order to design, train, and deploy the CNN+LSTM model, one uses the frameworks of Tensorflow and Keras.
- Data Processing: Data processing of any videos and images is done by openCV, and it gives options to manage the data efficiently through NumPy and Pandas. Scikit-learn is applicable in the performance analysis and computing of metrics.
- Hardware: There is support of both CPU and GPU execution but the acceleration provided by a graphics card has a significant role to play in addressing the high-computation workloads of training and real-time inference.
- Demonstration: The functionality of the model presented in an intuitive and user-friendly system will be demonstrated as a lightweight and interactive web application created with Streamlit.

VII. Conclusion

IIn conclusion, the issue of face-swap deepfakes has been confronted in this project because a robust AI/ML-based detection system was created. Classical methods of detection have a hard time with generative models that are quickly evolving. This paper presents a hybrid CNN-LSTM platform, which presents a scalable and novel platform by examining temporal inconsistencies and spatial anomalies in video data. The model is trained on benchmarks including FaceForensics++, thus, it can be anticipated that the model will attain high accuracy (above 95) and will also be resistant to novel deepfake methods. The strategy suggests speed, accuracy and efficiency. This study helps in promoting ethical AI, as well as further developing a more open and trustful media landscape in the era of synthetic media, through integration with wider digital ecosystems.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in IEEE International Workshop on Information Forensics and Security, 2018.

[2] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme viseme mismatches," in CVPR Workshop on Media Forensics, pp. 6603-661, 2020.

[3] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in Workshop on Media Forensics at CVPR, vol. 1, 2019. K. Elissa, “Title of paper if known,” unpublished.

[4] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based cnn," in Workshop on Media Forensics at CVPR, 2019.

[5] M. A. Rahman, S. K. Patel, and L. Chen, "CNN based Deep Learning model for Deepfake Detection," IEEE Transactions on Image Processing, vol. 30, no. 4, pp. 2845–2857, Apr. 2022.

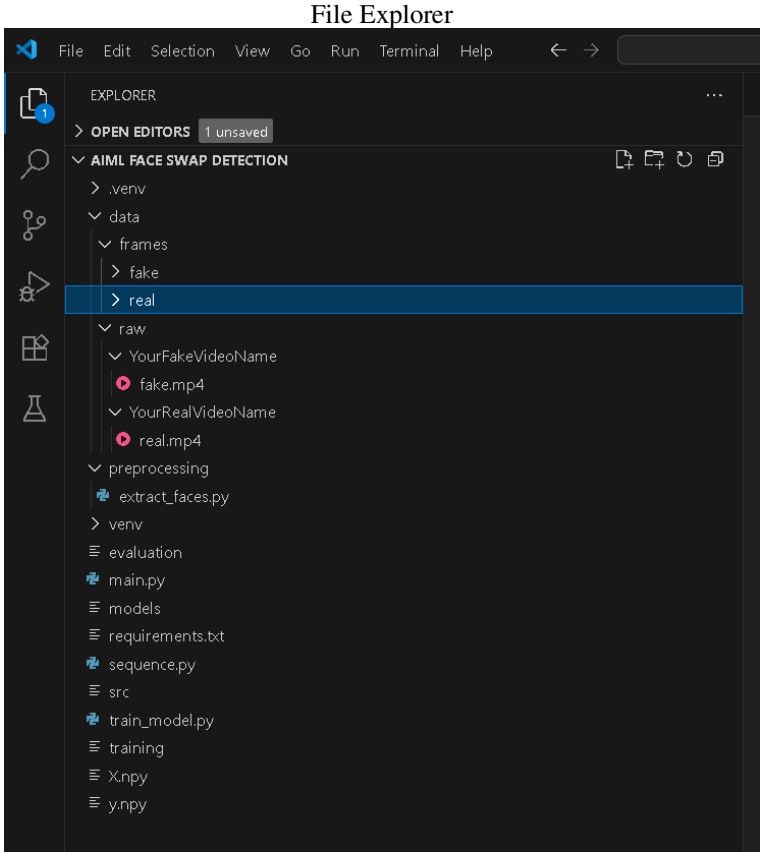
[6] K. J. Thompson, R. Singh, and A. Martinez, "Deepfake Video Detection Using Recurrent Neural Networks," IEEE Access, vol. 10, pp. 45123–45135, May 2022.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

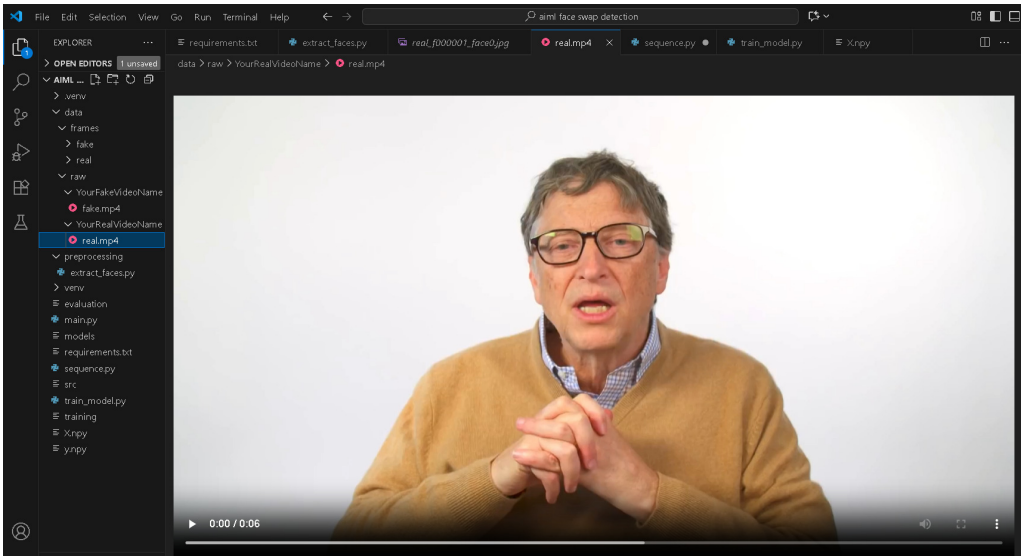
[8] M. Bohacek and H. Farid, "Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms," Proceedings of the National Academy of Sciences, vol. 119, no. 48, e2216035119, 2022.

DeepFake Testing Images (For Reference)

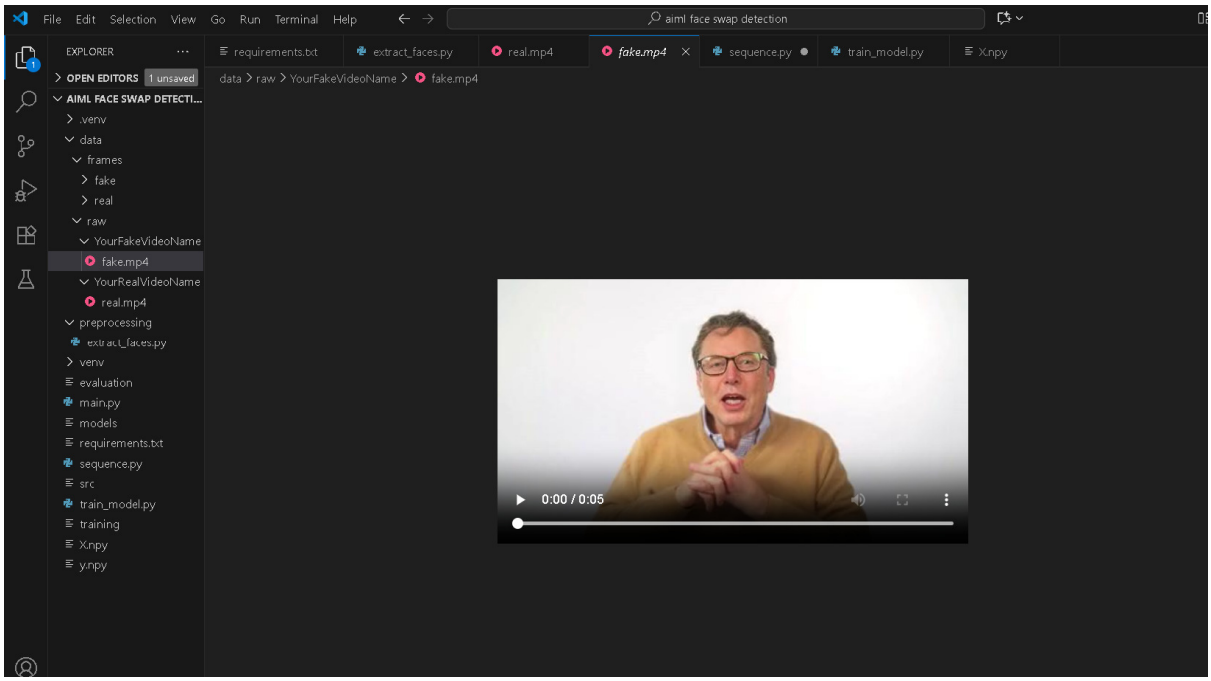
File Explorer



Real MP4 Video

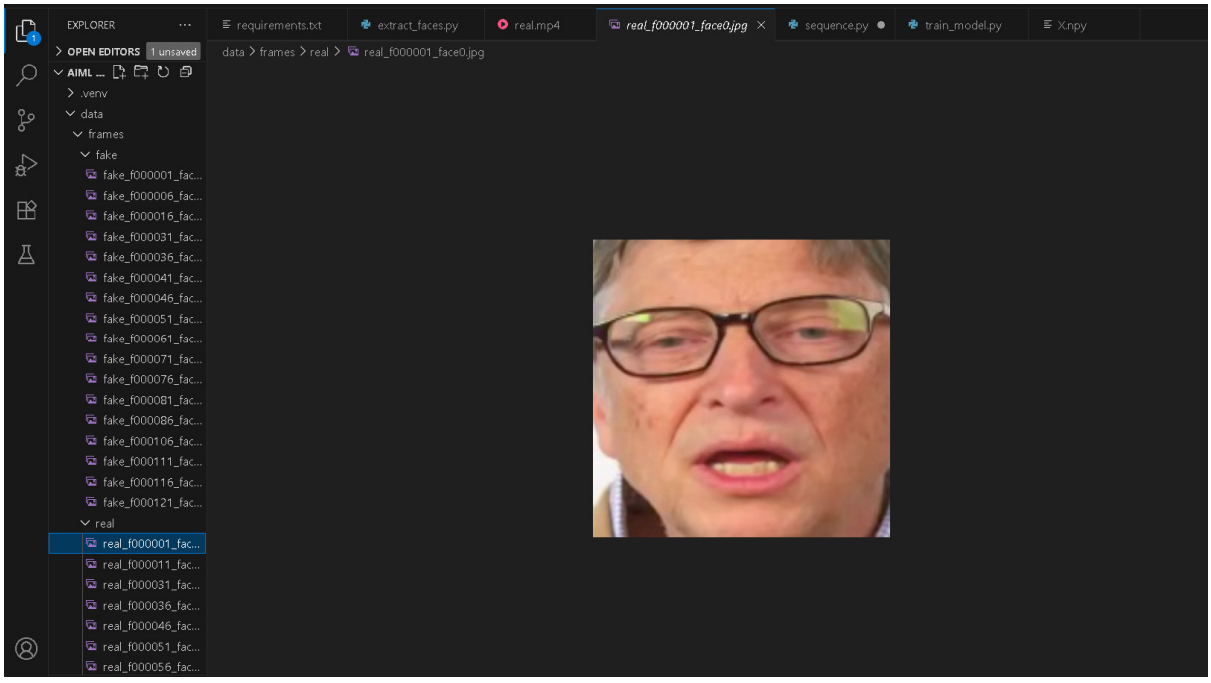


Fake MP4 Video



Preprocessing/Sampling

Real Frames:



Fake Frames: