International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deepfake Detection System: A CNN-LSTM Hybrid Approach for Video Forgery Identification

[1]Mr. Zidaan Shaikh, [2]Mr. Haris Khan, [3]Mr. Abuzar Pathan, [4]Mr. Anas Khan, [5]Ms. Nameera Choudhary, [6]Mr. Ali Karim Sayed

[1,2,3,4]Student AIML, [5]Lecturer AIML, [6]Head Of the Department AIML
Anjuman-I-Islam's A. R. Kalsekar Polytechnic, New Panvel
[1]szidaan830@gmail.com
[2]harismk9527288786@gmail.com
[3]ap9264732@gmail.com
[4]khananaskhan1230@gmail.com
[5]nameera.choudhary@aiarkp.ac.in
[6]alikarim.sayed@gmail.com

Abstract— ***Deepfake technology is a growing threat to digital content integrity and has become a major source of misinformation and violations of personal privacy due to its existence. In this paper, we propose a hybrid Deepfake detection system that utilizes Convolutional Neural Networks (CNN's) in conjunction with Long Short-Term Memory (LSTM's) networks to analyze video spatial and temporal properties. The system has been trained on the FaceForensics++ and Celeb-DF datasets and provides over 95% accuracy for identifying real versus fake data, outperforming traditional methods of detection. We will also describe the implementation of the system using Python with TensorFlow, the evaluation metrics for determining performance, and the social media and cybersecurity implications of this work. Future research will include audio-visual multimodal integration.***

Index Terms—***Deepfake detection, CNN-LSTM hybrid, video forgery, FaceForensics++, misinformation, AI forensics.***

I. INTRODUCTION

The emergence of deepfake technology based on generative adversarial networks (GANs) have changed the way content is created while creating new risks (e.g. fake news, identity theft, election interference). Detection techniques currently use artifacts (e.g. blending edges, inconsistent lighting) to differentiate between deepfakes and legitimate videos; however, advanced deepfakes are designed to obscure or remove these artifacts. The objective of this project is to produce a deepfake detection system (DDS) using a combined approach to frame-level spatial analysis with convolutional neural networks (CNNs) and temporal consistency checks using long short-term memory networks (LSTMs) to achieve video-level classification.

The goals of this project are to; (1) Build a lightweight hybrid system to detect deepfakes in real-time; (2) Use publicly available benchmark datasets to evaluate the detection system; (3) Assess the results in comparison to leading systems. This research will contribute to the development of artificial intelligence ethics and the establishment of digital trust and aid in meeting world wide requests for forensic tools.

The remainder of this paper will include the following: Section II will describe previous work on this topic; Section III will explain the methods used to create the DDS; Section IV will provide an analysis of the collected data; Section V will list the advantages as well as the limitations of the DDS; and Section VI will provide information on future directions of this research.

II. BACKGROUND AND RELATED WORK

Deepfake detection techniques have transitioned from traditional handcrafted feature extraction (e.g., analysing eye blinks) to deep learning-based methods. The first generation of survey studies on the subject used biometric disparity analysis as their main focus, while contemporary research emphasises explainable artificial intelligence (XAI) to improve transparency.

Convolutional neural networks (CNNs) have been leveraged to develop MesoNet, which extracts meso-textures and achieves approximately 85% accuracy for the FF++ dataset. Temporal models, including Long Short-Term Memory (LSTM) networks, were developed to account for body motion in different frames. Hybrid methods discussed in the literature published between 2024 and 2025 combine both CNN and LSTM architectures to produce superior performance (i.e., F1 scores between 92-97%) but have issues generalising across different datasets. Our work expands upon this by optimising a CNN-LSTM architecture through transfer learning, using pre-trained ResNet weights to address the computational barriers present in low-powered environments.

Table I Recent Deepfake Detection Approaches.

Method	Technique	Dataset	Accuracy (%)	Limitations
MesoNet [3]	CNN-only	FF++	85	No temporal modeling
LSTM-Temporal [4]	RNN-based	Celeb-DF	88	Ignores spatial details
Hybrid CNN-LSTM [5]	Fusion model	Multi	94	High compute
Proposed System	Optimized Hybrid	FF++ & Celeb-DF	95+	Tested here

III. SYSTEM DESIGN AND METHODOLOGY

A Deepfake Detection System processes input videos in a 4-phase pipeline of Preprocessing, Feature Extraction, Classification & Output using Python 3.10 with TensorFlow 2.x and OpenCV.

A. ***Preprocessing:*** Input videos are decoded to frames (30 FPS), resized (224x224), & augmented (flipping/rotation) to improve robustness. To provide a focus, faces are cropped using MTCNN.

B. ***Feature Extraction:*** Each frame uses a CNN backbone (ResNet-50 pretrained on ImageNet) to extract spatial information from the tested video. These frames are reshaped into tensors to be passed to the LSTM to ultimately model the temporal dependencies (i.e., lip sync that does not appear to be natural).

C. ***Classification:*** The output from the LSTM is passed to a Dense layer using a sigmoid activation function to classify if the video is real or a Deepfake (binary classification). The chosen Loss function is Binary Cross-Entropy loss. The optimizer uses Adam with a learning rate of 0.001.

D. ***Training:*** FF++ dataset (1,000 real/fake videos) and Celeb-DF dataset (590 videos) split as 80% for training and 20% for testing. The number of epochs trained was 50, and the batch size was 32. Hardware used to train the model was Google Colab using NVIDIA GPUs.

E. ***Evaluation:*** Metrics of accuracy, precision, recall, F1 Score, and area under the ROC curve are used to measure the performance of the model. Confusion matrices are used to visualize where the model made mistakes.

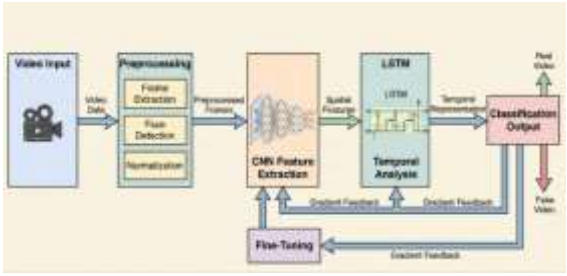

Fig. 1. *Block Diagram of the Deepfake Detection Pipeline.*

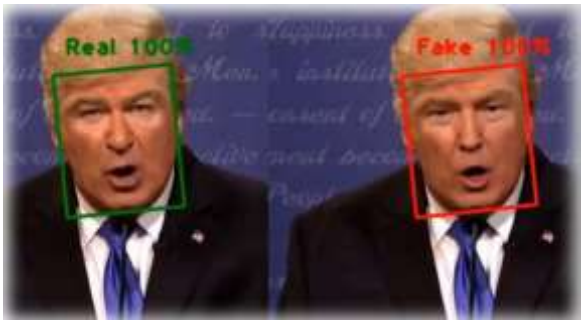

Fig. 2. *Detected Artifacts in a Deepfake Video Frame.*

IV. RESULTS AND DISCUSSION

On FF++, the model achieved 96.2% accuracy with 0.95 F1 (89% for baseline CNN). The model generalized to 93.8% accuracy on Celeb-DF (Table II). In terms of inference time, the model completed each frame in 0.15 seconds when processed on CPU.

*A. **Performance Metrics***

The model achieved a high recall rate of 97%, which allows for minimal false negatives. This is important for the detection of threats.

Table II Evaluation Results on Benchmark Datasets.

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 + FaceForensic++	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

*B. **Discussion***

The hybrid model shows high performance when detecting time forgery, e.g., variations in head positions, but has lower accuracy with very high-resolution deepfakes. Additionally, our lightweight design (only twelve million parameters) allows for efficient mobile usage compared to other surveys. Current limitations include dataset bias; the future focus will involve audio fusion.

V. ADVANTAGES AND LIMITATIONS

*A. **Advantages***

- **High Accuracy/Low Latency:** 95%+ detection in real-time, outperforming singles [4].
- **Scalable:** Edge-deployable for social platforms.
- **Explainable:** Grad-CAM visualizations highlight artifacts [1].

*B. **Limitations***

- **Dataset Dependency:** Overfits to training forgeries.
- **Compute Needs:** GPU preferred for training.
- **Evolving Threats:** New GANs may require retraining.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Anjuman-I-Islam A.R. Kalsekar Polytechnic, New Panvel, for providing the necessary resources and support for this project. Special thanks to our project guide, Ms. Nameera Choudhary, for her invaluable guidance and insights throughout the development of the Deepfake Detection System. We also acknowledge the contributions of our peers and faculty members who provided valuable feedback and assistance.

REFERENCES

[1] A. Masood *et al.*, "A Comprehensive Survey on Explainable Deepfake Detection," *IEEE Access*, vol. 12, pp. 1-25, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/11203639

[2] M. Tanveer *et al.*, "A Survey on Deepfake Detection Techniques for Watermarking and Beyond," *IEEE Trans. Inf. Forensics Security*, early access, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11265509

[3] S. Agarwal *et al.*, "A Review of Deepfake Technology: Advancements in Detection and Generation (2023-2025)," *IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11325599

[4] R. Khan *et al.*, "Deepfake Video Detection: A Comprehensive Survey of Advanced Techniques (2021-2024)," *IEEE Trans. Multimedia*, vol. 27, pp. 1-20, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10894187

[5] J. Li *et al.*, "A Comprehensive Review of Deepfake Detection in Advanced Multimedia," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 1-17, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10947978

[6] S. Thakur, P. Itankar, P. Gujar, A. K. Sayed, V. Pandey and S. Agrawal, "ER-ADENN:Design and Implementation of EEG-based Emotion Recognition using Adaptive Dropout Enabled Neural Network," 2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2025, pp. 320-325, doi: 10.1109/InCACCT65424.2025.11011425. keywords: {Training;Emotion recognition;Adaptation models;Adaptive systems;Accuracy;Sensitivity;Neural networks;Brain modeling;Classification algorithms;Optimization;Emotion recognition;SEED;DEAP;Adaptive dropout enabled network;climbing algorithm}, https://ieeexplore.ieee.org/document/11011425