IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

Detecting Deepfake Faces with CNN and LSTM Models

Prajwal CG[1], Priyanka HU[2], Rahul GP[3], Varshini MG[4]
Department of Computer Science and Engineering[1-4]
Kalpataru Institute of Technology, Tiptur, India

Abstract: *Trust in digital media is seriously jeopardized by deepfake videos made with sophisticated generative models, which enable realistic but fake content that is hard for people to recognize. In this paper, a hybrid ResNeXt-50 and Long Short Term Memory (LSTM) architecture is used to combine spatial and temporal cues in a video-based deepfake detection system that uses face regions extracted from each frame. The model is trained using preprocessed face-only clips from public benchmark datasets like FaceForensics++, Celeb-DF, and the Deepfake Detection Challenge, as well as a recent Kaggle deepfake video corpus. The suggested method is integrated into a Django web application that enables users to upload a video and receive a real/fake prediction in close to real time. It achieves promising accuracy and balanced precision–recall on held-out videos. These findings show that face-region preprocessing and joint CNN-LSTM modeling offer a practical and efficient way to detect deepfake videos.*

Keywords: Django, video forensics, LSTM, ResNeXt-50, deepfake detection, transfer learning

I. INTRODUCTION

Digital media has become one of the most common means of communication and information sharing across social networks, news platforms, and online services. In recent years, rapid progress in generative models such as Generative Adversarial Networks (GANs) and diffusion-based techniques has made it possible to create highly realistic manipulated videos, commonly referred to as deepfakes. These videos can convincingly alter a person's facial identity, expressions, or speech, making it extremely difficult for human observers to differentiate between authentic and manipulated content.

The widespread availability of deepfake technology poses significant risks to digital trust and security. Deepfake videos can be misused for identity impersonation, financial scams, political misinformation, and social manipulation. As the quality of generated videos continues to improve and as content is frequently compressed and shared on social media platforms, traditional forensic methods and manual inspection techniques have become increasingly ineffective.

To address this issue, automated deepfake detection techniques based on deep learning have gained considerable attention. Many existing approaches rely on convolutional neural networks to identify spatial artifacts within individual video frames. While effective to some extent, such methods often fail to capture temporal inconsistencies that occur across consecutive frames. In practice, deepfake videos often exhibit subtle temporal artifacts such as unnatural facial movements, inconsistent expressions, and frame-to-frame flickering.

In this paper, a video-based deepfake detection framework is proposed that jointly models spatial and temporal information using a hybrid ResNeXt-50 and Long Short-Term Memory (LSTM) architecture. The proposed system focuses exclusively on facial regions extracted from video frames in order to minimize background interference and emphasize manipulation-related artifacts. The model is trained and evaluated on multiple publicly available benchmark datasets, including FaceForensics++, Celeb-DF, the Deepfake Detection Challenge (DFDC), and a recent Kaggle DeepFake Videos 2025 dataset. In addition, the trained model is deployed through a Django-based web application, enabling users to upload videos and obtain real-time predictions, thereby demonstrating the practical applicability of the proposed approach.

DOI: 10.48175/IJARSCT-30746

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

II. LITERATURE SURVEY

Early methods for detecting deepfakes relied on manually created characteristics like erratic eye blinking, inconsistent head posture, or colour differences between the face and background. These approaches tend to fail as synthesis models get better, but they are effective for certain generation techniques. Convolutional neural networks (CNNs) trained on altered images and frames became the predominant method with the advent of deep learning, producing impressive results on datasets like FaceForensics++ and Celeb-DF.

Nevertheless, a lot of these approaches disregard temporal coherence because they handle frames separately. Several studies apply 3D CNNs or recurrent neural networks to frame sequences in order to incorporate temporal information. Improved robustness on video benchmarks has resulted from the use of recurrent models based on Long Short-Term Memory (LSTM) or gated recurrent units to identify temporal flickering and variations in lighting and facial expressions across frames. More recent studies have looked into cross-dataset training techniques to enhance generalization as well as transformer architectures and attention mechanisms for modeling long-range temporal dependencies.

The gap between deployment-ready systems and academic benchmarks persists despite these advancements. Many studies stop at offline evaluation without developing an end-to-end application, evaluate only on one dataset, or ignore contemporary AI-generated video datasets. The work described in this paper fills these gaps by integrating the model into a useful web-based tool and combining multi-dataset training with a robust CNN–LSTM backbone.

III. PROPOSED SYSTEM

A. Overall Architecture

The suggested system uses a modular pipeline that starts with the upload of a video and concludes with a binary prediction that indicates whether the input is authentic or fraudulent. The Django web interface allows users to upload videos, which are then stored by the backend and sent to the preprocessing module. This module creates standardized face-only clips, extracts frames, and finds faces. After that, the trained ResNeXt-50 + LSTM model produces a probability score for every class. Lastly, the user is shown the prediction and confidence score along with some basic logging data on the web interface.

B. System Modules

The main system modules are:

• Data management: scripts and a structure for storing and retrieving videos from Kaggle DeepFake Videos, Celeb-DF, DFDC, and FaceForensics++.

• Preprocessing service: a Python/OpenCV pipeline that crops face regions, detects and aligns faces, decodes videos, and saves them as fixed-size clips.

• Model training module: Jupyter/Colab notebooks that use PyTorch with GPU acceleration to implement dataset loading, model definition, training, and evaluation.

• Inference API and web front-end: Django views and templates that expose model inference as a REST-like endpoint and provide an HTML interface for uploading and visualizing results.

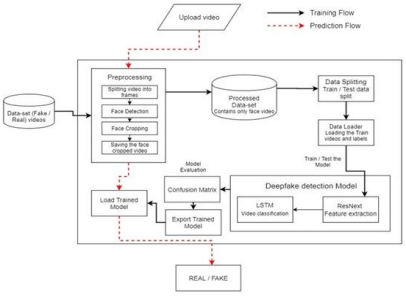

Fig. 1. Overall system architecture: video ingestion, preprocessing, model inference and web interface.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

IV. METHODOLOGY

A. Datasets

This work makes use of four publicly available datasets. Real and altered videos created using various facial reenactment and replacement techniques can be found in FaceForensics++. High-quality celebrity deepfakes that are more difficult to spot visually are the focus of Celeb-DF. Similar to social media, the DFDC dataset offers a vast and varied collection of crowdsourced real and fake videos. Lastly, the Kaggle DeepFake Videos 2025 dataset helps assess generalization to newer generation techniques and includes more recent manipulations.

B. Preprocessing Pipeline

OpenCV is first used to decode each video at a predetermined frame rate. The primary face region is located for each frame using a face detector based on the face_recognition library and dlib landmarks. The ImageNet mean and standard deviation are used to normalize the detected faces after they have been cropped and resized to 112×112 pixels. A predetermined number of frames are uniformly sampled from each video to create an input sequence; videos with insufficient valid face frames are eliminated. In order to help the network concentrate on subtle facial artefacts rather than background variations, the learning model uses the resulting face-only clips as input.

C. Model Architecture

A ResNeXt-50 (32×4d) network pre-trained on ImageNet is used for spatial feature extraction. After global average pooling, the residual convolutional backbone of ResNeXt-50 generates a 2048-dimensional feature vector for every input frame once the final classification layer is removed. An LSTM network with a hidden size of 2048 receives the temporal sequence created by these frame-level features. To determine the likelihood that the video is authentic or fraudulent, the final hidden state of the LSTM is passed through a fully connected layer with dropout and a softmax activation. The model is able to capture both temporal inconsistencies and local spatial artefacts that are characteristic of deepfake videos thanks to its hybrid design.

D. Training Setup

While making sure that identities do not overlap between splits, videos from all datasets are combined and randomly divided into training, validation, and test subsets. The model is implemented in PyTorch and trained using the Adam optimizer on an NVIDIA GPU with a weight decay of 10^{-5} and an initial learning rate of 5×10^{-5}. Due to memory limitations, a mini-batch size of four videos is utilized. To prevent overfitting, the network is trained for a maximum of 25 epochs with early stopping based on validation loss. To increase robustness during training, basic data augmentations like random horizontal flipping and mild colour jitter are applied at the frame level.

E. Evaluation Metrics

Standard binary classification metrics are used to assess performance. The overall percentage of videos that are correctly classified is known as accuracy. The "fake" class is considered positive when calculating precision, recall, and F1-score, which indicates how effective the detector is at identifying manipulated content. To evaluate the trade-off between true and false positive rates, the area under the receiver operating characteristic curve (AUC) is also provided. To see the distribution of true positives, true negatives, false positives, and false negatives for each experiment, confusion matrices are plotted.

V. RESULTS AND DISCUSSION

A. Quantitative Results

The performance of the suggested model on the distinct test sets generated from each dataset is compiled. With only a slight decline on the more difficult Kaggle 2025 videos, the results show that the hybrid ResNeXt-50 + LSTM architecture maintains high accuracy and F1-score across all datasets. The classifier reliably distinguishes between real and fake samples across a broad range of thresholds, as evidenced by the AUC values near 1.0.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, January 2026
Impact Factor: 7.67

B. Confusion Matrix and Analysis

The confusion matrix for the combined test set shows that the majority of videos are correctly classified, as indicated by the large values along the diagonal. A tiny percentage of authentic videos—usually low-light or highly compressed clips with diminished facial details—are mistakenly labeled as fraudulent. Occasionally, some challenging fake videos with nuanced facial expressions and excellent rendering are predicted to be authentic. These findings imply that adding attention mechanisms or higher resolution crops could further lower the residual errors.

C. Ablation Study

To separate the effects of temporal modeling and extra training data, an ablation study was carried out. A baseline model that solely uses the ResNeXt-50 backbone with average pooling over frames achieves a lower F1-score, suggesting that handling borderline cases is challenging. Performance is enhanced on all datasets, especially those with complex motion, when the LSTM layer is added. Training without the Kaggle 2025 videos results in discernible decreases in accuracy, demonstrating the significance of incorporating recent manipulations for strong generalization.

VI. CONCLUSION

This study introduced a deepfake video detection system that uses face-region clips and a hybrid ResNeXt-50 and LSTM architecture. In order to extract and normalize faces from various public datasets, a unified preprocessing pipeline was created. The resulting model performed well on a number of difficult benchmarks. The usefulness of the method was demonstrated by integrating the trained network into a Django web application that lets users upload videos and receive real/fake predictions.

Future work will investigate transformer-based architectures for longer temporal modeling, perform continuous learning as new deepfake generators and datasets become available, and expand the system to multimodal detection by integrating audio signals and lip-sync analysis. Other areas of interest include optimizing inference speed for large-scale deployment and testing the system on genuinely real-world social media content.

ACKNOWLEDGMENT

The authors would like to thank their project supervisor and the Computer Science Department at Kalpataru Institute of Technology for guidance and computational resources.

REFERENCES

[1] O. Rössler et al., “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. ICCV, 2019.
[2] Y. Li et al., “Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics,” in Proc. CVPR, 2020.
[3] B. Dolhansky et al., “The Deepfake Detection Challenge Dataset,” arXiv:2006.07397, 2020.
[4] M. Karki, “DeepFake Videos Dataset,” Kaggle, 2025.