iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** VIII **Month of publication:** August 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.84766

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com*

# Detection of Face-Swap Based Deepfake Videos Using Hybrid CNN-LSTM Architecture

Suraj S. Pawar[1], Kaustubh R. Saswade[2], Nikhil R. Mane[3], Jay L. Borate[4], Anurag S. Karande[5]
*Department of Computer Science, PES'S College of Engineering Phaltan, Tal-Phaltan, Dist-SATARA, India*

***Abstract:*** ***The development of deepfake technologies due to breakthroughs in AI and deep learning allows producing highly realistic manipulated videos and audio, thus posing a threat to misinformation and digital security. Despite deepfake technology having several legitimate uses, including use in the media industry, its inappropriate use for distributing fake news, impersonation, and cyber attacks demands the development of reliable detection techniques. The current state-of-the-art approaches of detecting deepfakes mostly utilize spatial or temporal analysis based on CNNs and RNNs; however, most existing approaches fail to generalize well and are unable to recognize more complicated manipulations on a variety of different data sets. This paper aims at developing a novel multimodal approach to deepfake detection, integrating spatial, temporal, and audio features. The proposed system makes use of CNN-based architectures to extract spatial information from the input images, transformers to capture the temporal information, and Mel-frequency cepstral coefficients (MFCC) to analyze the audio data. These heterogeneous features are combined using attention-based learning to improve the classification accuracy. The proposed method was tested on various benchmarking datasets, including FaceForensics++, DeepFake Detection Challenge (DFDC), and Celeb-DF, yielding higher accuracy than current methods. Experimental results indicate that the combination of multimodal features enhances the detection capacity. The proposed model is highly efficient in addressing deepfake challenges in the digital forensic field.***

## I. INTRODUCTION

The digital media landscape has changed a lot because artificial intelligence is moving so quickly and powerful computers are becoming more common. Deepfake technology is one of the most important new things in this field. It uses deep learning techniques especially generative adversarial networks(GANs),to make videos and audio that look very real but are actually fake. Deepfakes change or replace facial expressions, identities or speech in multimedia content which makes it hard to tell the difference between real and fake data. This technology is useful for making movies, virtual simulations and protecting privacy but when it is misused , it raises serious ethical, social, and security issues. The rise of deepfake content on social media has caused false information, political manipulation, identity theft, and a loss of trust in digital content. Most traditional ways to find deepfakes have used convolutional neural networks to look for visual inconsistencies in individual frames. These methods work well for finding some types of manipulations, but they often miss temporal inconsistencies between video frames, which are very important signs of deepfake generation. To overcome this limitation, researchers have implemented recurrent neural networks, including long short-term memory (LSTM) models, for the analysis of sequential frame data.

But these models have problems with long-range dependencies and need a lot of computing power. Also, most current methods only use visual features and ignore the audio part, which can also have important clues for finding manipulated content.Recent progress in deep learning, especially transformer-based architectures, has been very successful at finding long-range dependencies and making sequential data analysis tasks work better. The incorporation of multimodel data, encompassisng visual and auditory elements, has surfaced as a viable strategy for improving deepfake detection system. This paper proposes a multimodel deepfake detection framework that integrates spatial, temporal, and audio analysis through sophisticated deep learning methodologies, motivated by recent advancements. The suggested method seeks to enhance detection precision, resillence, and generalization across various datasets, thereby overcoming the shortcomings of current techniques and adding in the advancement of more dependable digital media authentication systems.

## II. LITERATURE REVIEW

Because synthetic media generation techniques are getting better, deepfake detection has become an important area of research. Initial methodologies in this field predominantly concentrated on detecting visual artifacts and discrepancies arising during the deepfake creation process. For this purpose, convolutional neural networks (CNNs) have been widely used because they can find spatial features in images and spot problems like color mismatches, blending artifacts, and textures that don't look natural. Models like XceptionNet and MesoNet have shown that they can find fake images and videos very well. These methods, on the other hand, are only good for analyzing frames and often miss temporal inconsistencies between video sequences.

To overcome the constraints of spatial-only models, researchers have integrated temporal analysis methodologies utilizing recurrent neural networks (RNNs), specifically long short-term memory (LSTM) networks. These models look at sequences of frames to find inconsistencies in facial movements, expressions, and transitions. CNN-LSTM hybrid models have made detection performance better, but they still have trouble with long-range dependencies and need a lot of computing power. Also, their performance often gets worse when they are tested on datasets that are different from the ones they were trained on, which shows how bad generalization is.

Recent research has investigated the application of capsule networks and transformer-based architectures for the detection of deepfakes. Capsule networks try to keep the hierarchical relationships between features, which makes them more resistant to complex changes. On the other hand, transformer-based models use attention mechanisms to find global dependencies in data, which makes them very good for video analysis tasks. Moreover, numerous researchers have suggested multimodal methodologies that integrate visual and auditory attributes to improve detection precision. These methods use audio problems, like lip sync that doesn't match and speech patterns that don't sound natural, to find altered content.

Even with these improvements, problems like computational complexity, the absence of standard evaluation benchmarks, and the need for robustness across different datasets are still not solved. So, we need a complete and effective deepfake detection system that uses multiple modalities and advanced deep learning methods to make it work better and more reliably.

## III. SYSTEM ARCHITECTURE

The proposed system for deepfake detection can be modeled as a pipeline-based approach in which the input videos go through various phases, which include pre-processing, feature extraction, model training, and prediction. The architecture is capable of supporting both training flow and prediction flow, as shown in Figure, allowing the algorithm to learn from a set of labeled data and make predictions for unknown videos.This whole process starts from the input data, which includes real and deepfake videos. All the videos go through the pre-processing phase, which plays an important part in making the videos ready for training the model. During this phase, each video clip is divided into its constituent frames, thus translating the temporal video input data to an image sequence. After extracting the frames, face detection techniques are utilized to detect the faces in the frames using algorithms like Multi-task Cascaded Convolutional Networks (MTCNN) or Haar cascade. After identifying the faces, face cropping is done to crop out the necessary portion of the face from the frames, eliminating unnecessary background data. Once preprocessed, the data is saved and sent to the data splitting stage, where it is separated into training and testing data. The use of unseen data improves generalization capacity of the model, which is critical when detecting deepfake videos. The data loader module takes care of the task of loading the training set and respective labels in batches to improve the performance of the model.

The most important stage of the project is the model used to detect deepfake videos. The model contains two parts: feature extraction and video classification. In order to extract the features from each frame, a ResNeXt CNN is used. The model detects high-level features, including artifacts, inconsistencies in color and texture, and others, which may be considered to be suspicious in the context of deepfake creation. The next step is the video classification, during which an LSTM network processes temporal information of the frames and recognizes possible inconsistencies in face movement and expression.

During training, the model learns to distinguish between real and fake videos by reducing classification errors. Model evaluation is done by analyzing a confusion matrix that reveals information regarding accuracy, precision, recall, and misclassifications. When the model performs satisfactorily, it is stored and exported for future use.

The prediction process involves uploading a new input video by the user. It goes through the same preprocessing, involving extraction of frames, detecting faces, and cropping. After processing, the frames are fed to the trained model, which uses ResNet and LSTM to learn spatial features and temporal sequences. The result is a classification output showing either REAL or FAKE and its level of confidence.

In conclusion, the architecture is capable of capturing spatial features and modeling temporal sequences to detect deepfakes. CNN and LSTM ensure high accuracy due to their high-level abstraction capacity. Additionally, the architecture is scalable and can accommodate future improvements like adding an audio component or transformer model.

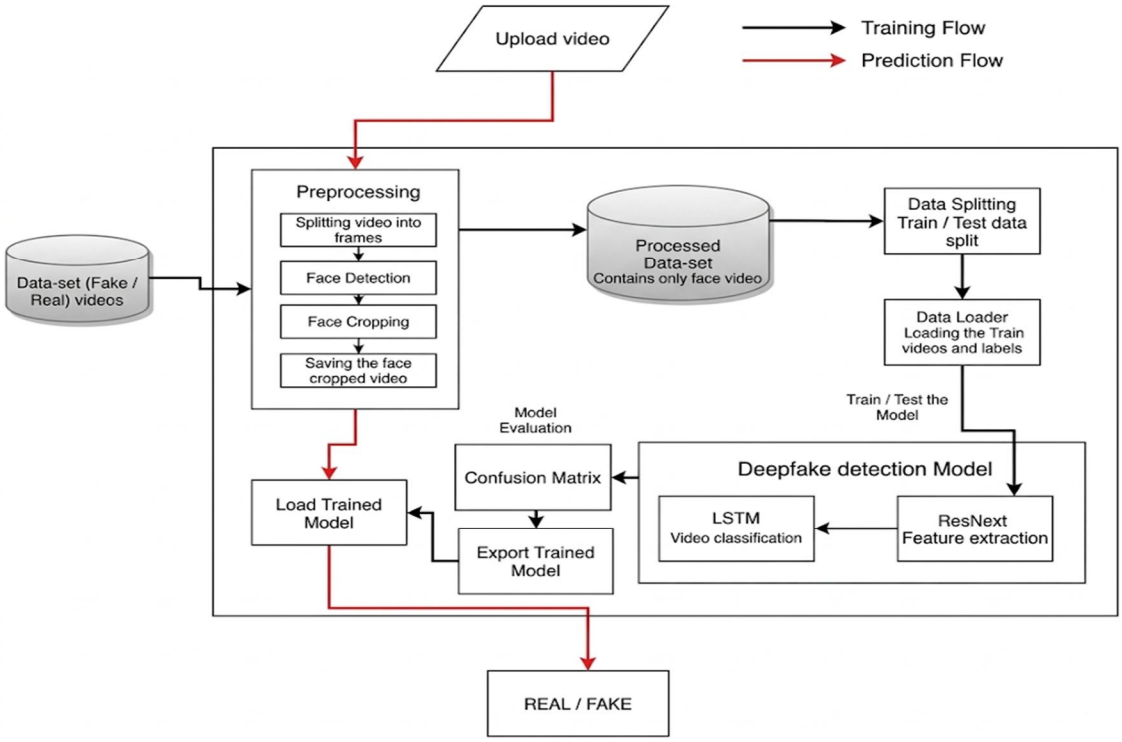



**Figure 1: System Architecture**

## IV. METHODOLOGY

The proposed deepfake video detection method uses a systematic approach consisting of several steps, including data collection, data preprocessing, feature extraction, multimodal feature fusion, model training, and model evaluation. The purpose of this methodology is to detect deepfake videos efficiently by exploiting the spatial, temporal, and audio properties of videos through deep learning. The first step involves data collection, which consists of collecting a variety of deepfake and real videos from benchmark datasets such as FaceForensics++ (FF++), DeepFake Detection Challenge (DFDC), and Celeb-DF. All the benchmark datasets include a vast number of deepfake videos that have been created using different deepfake generation methods. This ensures that the algorithm generalizes well on any type of deepfake video and does not depend only on the benchmark dataset.

During the pre-processing phase, videos are first broken down into frames at a constant frame rate in order to retain temporal information while making computations less complex. Faces from each of the video frames are detected using multi-task cascaded convolutional networks (MTCNN) and other similar face detection algorithms. The faces are then cropped to a fixed size in order to keep uniformity within the whole data set. Further pre-processing steps include normalizing images, removing noise, and augmenting the data through flipping, rotating, and adjusting image brightness levels.

After preprocessing, feature extraction is performed with the use of multiple modalities. The spatial feature extraction involves using a convolutional neural network like EfficientNet or ResNeXt to detect the fine details associated with video manipulation, which include inconsistencies, distortions, and blurring. Temporal relationship modeling involves passing the frame-wise extracted features into a transformer network that takes care of the relationship between various frames and detects inconsistencies with regard to the movements and expressions in the face. This method works better than the recurrent neural networks because of the handling of long-term dependencies and parallel processing capabilities.

Besides the visual analysis, the methodology also uses audio features to analyze and detect inconsistencies in the way people speak in deepfake videos. An audio signal will be extracted from the video and transformed into mel frequency cepstral coefficient (MFCC) to represent the spectral properties of audio. These extracted features will then be analyzed by a convolutional neural network to detect any anomalies such as mismatch of lips synchronization and unnatural speech properties. Finally, after extracting the necessary spatial and temporal features as well as audio features, all of them will be integrated in the feature fusion stage of the methodology. In this stage, attention-based feature fusion technique will be used to ensure the right significance of each feature is assigned.

During the model training process, the merged feature vector is fed to fully connected layers in order to classify the images into two categories: deepfake or real. During the training process, the model is optimized based on the binary cross-entropy loss function and trained by using the Adam optimizer, along with an appropriate learning rate to obtain optimal results. In addition, regularization techniques such as dropout can be used during the training process to avoid overfitting.

Finally, the performance of the model can be evaluated using traditional performance evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The metrics mentioned above will be used to evaluate the performance of the model in terms of its effectiveness in detecting deepfakes and distinguishing them from real videos. The process will be carried out on the testing dataset that includes unseen samples.

## V. RESULTS AND DISCUSSION

The performance of the suggested CNN-LSTM model is measured by common measures like accuracy, precision, recall, and F1-score. Performance analysis is done by creating a confusion matrix that can classify predictions into true positive, true negative, false positive, and false negative categories.

| | Predicted Real | Predicted Fake |
|---|---|---|
| Actual Real | TP | FN |
| Actual Fake | FP | TN |

The performance metrics are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

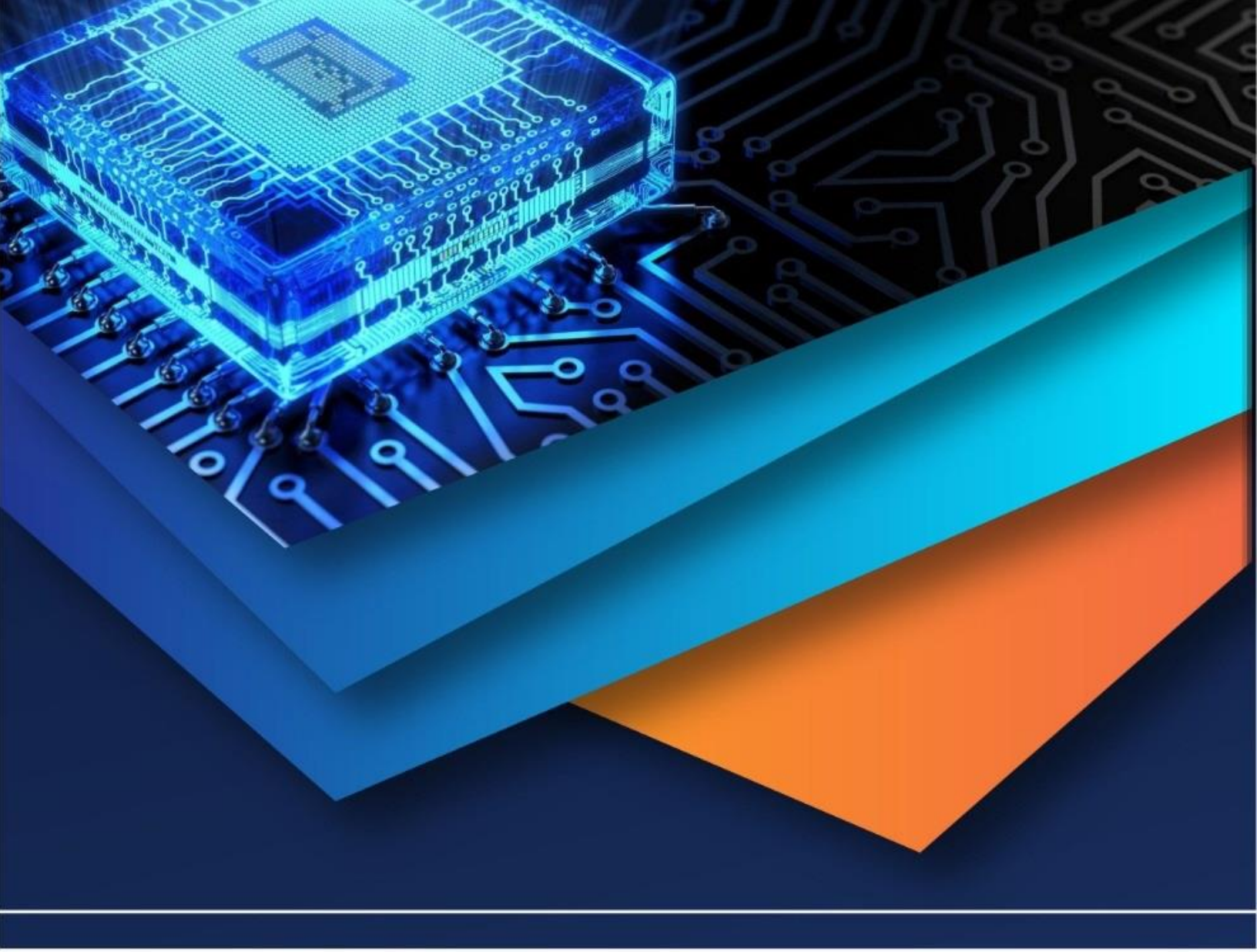
The findings show that the hybrid architecture has a higher performance than the CNN architecture and the LSTM architecture since it manages to efficiently address the inconsistencies both spatially and temporally. This architecture performs with high precision and consistency when used in different types of data sets. Nevertheless, there are weaknesses in this technique, such as its computational complexity.

## VI. CONCLUSION

In this paper, we present a combined CNN-LSTM model that can detect face-swap deepfake videos using spatial and temporal analysis techniques. The proposed approach demonstrates superior accuracy in identifying the manipulated videos and performs better than conventional approaches. The combination of CNN and LSTM models makes the proposed detection method more accurate and robust. It is evident from the paper that there is a need for development of advanced detection techniques to deal with the challenges of deepfake technologies.

## REFERENCES

[1] Mulani, C. R., et al., “Development of AI/ML Based Solution for Detection of Face Swap Based Deepfake Videos,” International Journal for Multidisciplinary Research, 2025.
[2] Malik, A., et al., “DeepFake Detection for Human Face Images and Videos: A Survey,” IEEE Access, 2022.
[3] Sunkari, V., et al., “Artificial Intelligence for Deepfake Detection: Systematic Review and Impact Analysis,” IAES International Journal of Artificial Intelligence, 2024.
[4] Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3
[5] Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.
[6] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.
[7] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv:1901.02212v2.
[8] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2
[9] I an Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.
[10] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY