International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Temporal and Spatial Feature Analysis for Deepfake Video Detection Using Recurrent and Convolutional Neural Networks

Dr. Bhanu Prakash Battula[1], D. Saniya[2], D. Devi Lakshmi[2], A. Anusri[2], D. Priyavallika[2]
[1]*Professor & Head,* [2]*B.Tech Students, Department of CSD, KKR & KSR Institute of Technology and Sciences, Guntur, India*

***Abstract*—The fast growth and development of artificial in- telligence, machine learning, and deep learning technologies, resulting in new technologies being used in the manipulation of multimedia and other data types. Although these technologies have been applied correctly in, for instance, the entertainment and educational fields, they have, however, been inappropriately and wrongly applied, especially in the production of very high- quality manipulated multimedia, now known as Deepfake, for the purpose of spreading misinformation as well as carrying out harmful activities such as harassments and blackmails. Multimodal detection techniques are used in deepfake detection for the purpose of determining whether or not a video, audio, as well as image, has been manipulated or generated artificially [1], [2]. We will conceptualize the creation of a deep learning model that will enable the detection of deepfake videos, hence addressing some real-world problems arising from deepfake technologies [1]. Our project provides a spatio-temporal deepfake detection framework: a unified approach that analyzes both video content and noise artifacts introduced during the process of deepfake generation. The model uses CNN to analyze individual frames for finding various spatial inconsistencies like blending errors or texture irregularities while using RNN with LSTM in order to model temporal dependencies across consecutive frames. By learning both spatial noise patterns and temporal inconsistencies together, this proposed CNN-LSTM model improves the reliabil- ity of deepfake detection. Unlike previous models, which either analyze spatial or temporal, not both, this combined approach will turn out to be more effective in identifying manipulated videos. These results demonstrate that applying both spatial and temporal analysis can substantially reinforce deepfake detection performance. Thus, the proposed system is well-suited for real- world applications such as digital forensics, media verification, or content authentication.**

***Index Terms*—Deepfake Detection, Convolutional Neural Net- work (CNN), Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Spatial Feature Extraction, Temporal Feature Analysis.**

I. Introduction

Technologies for editing images and videos are continuously evolving. There has been increased innovation in terms of image modification as well as video recordings.

Currently, it is achievable to create highly realistic and advanced images using fewer resources and simplified tutorials easily available on the internet. Deepfake is a technology used to replace the image or identity of a person in a video with that of another person in a video [1], [2]. In simpler words, it involves combining a person's image in terms of their features in an already pre-existing video to produce a highly manipulated visual replica. Although such technologies have profound advantages in terms of entertainment and mass media, they however pose significant hurdles as well. Deepfakes in videos can be used to create manipulated versions of videos to spread misinformation, as well as undermining reputations in terms of illicitly influencing public views. More specifically, it becomes challenging for people to determine if a video has been manipulated by simply viewing it if deepfakes are used. As a result of the growing realism in video deep fakes, manual verification is not considered reliable any longer. Humans may not be able to detect minute anomalies in a video [1], [6]. This further reinforces the need to depend on automated verification tools that verify the content in a far more detailed manner [4].

In our project, we concentrate on identifying manipulation in video utilizing deep learning methods. In the case of video, the system checks discrepancies in facial area information for each frame as well as activities in a sequence of frames. Through a thorough analysis of images our system has a more accurate method for identifying deepfakes.

The key goal of this project work is the improvement of authenticity and integrity of digital media. The proposed system can be applied in such areas as digital forensic analysis, media analysis, cyber security, and monitoring of social media where the identification of counterfeited video and audio clips is very paramount.

II. Problem Statement

The rapid development of AI-driven manipulation tools has enabled the creation of highly realistic "Deepfakes" that are becoming ever more indistinguishable from real media.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Classic detection methods tend to focus on either spatial artifacts, such as image-based noise, or temporal inconsistencies, including frame-to-frame jitters, but not both. As such, these single-modality models tend to miss high-quality deep fakes where spatial anomalies have been smoothed out or temporal transitions have been artificially stabilized. There is an imperative need for a unified framework that can collectively analyze intra-frame spatial irregularities and inter-frame temporal dependencies toward robust media authentication in digital forensics.

III. Literature Review

As AI is continuously evolving at a rapid pace, technologies such as deepfakes are also getting more refined. Many research studies are being carried out on analyzing how deepfakes can be created in videos. The early stages of research work on such technologies primarily focused on detecting the markers found within manipulated image or video files. The techniques applied within these early studies primarily detected frames within videos containing discrepancies in facial details, illumination effects, or portion blending [3], [4]. This work appeared to provide promising outcomes for low-quality deepfakes but could not deliver effective results in cases where higher-quality fake videos were considered.

The subsequent research introduced techniques from deep learning, particularly Convolutional Neural Networks (CNNs), which helped improve the detection results. CNNs examine spatial information from video, images and learn characteris- tics that distinguish real faces from fake ones [4]. The authors demonstrated CNN capability to capture information about abnormal texture and warped facial features. Nevertheless, almost all CNN-based models fail to notice inconsistencies between consecutive images and instead examine each image separately [8]. Therefore, these models overlook inconsisten- cies evident when images are reviewed over time together.

However, due to the constraint in the previous method, some studies have explored the applicability of Recurrent Neural Networks (RNNs), particularly Long Short-Term Mem- ory (LSTMs). These types of neural networks are generally developed for the patterns that appear in video sequences. It has been shown that LSTMs are quite effective for recognizing any abnormal facial expressions, blinking, and changes in facial expressions [5], [7]. For this purpose, the task of spatial feature extraction from the input image is performed by the CNNs.

Based on the literature reviewed, it has been seen that despite considerable advancements in the area of deepfake detection, some limitations exist concerning the handling of real-world inputs. In fact, several methods target a specific form or type, resulting in low robustness. This indicates the necessity for a holistic solution that emphasizes spatial, temporal, as well as audio-driven analysis [1], [6]. Noting this, the research project proposed here seeks to create a holistic deepfake detection model to provide a remedy to the limitations associated with earlier models.

IV. Proposed System

The architecture of the suggested system is aimed at addressing the dual aspect of deepfake manipulations, which are both frame-based, in terms of space, focusing on the presence of artifacts in individual frames, as well as a time-based issue, since a sequence of videos is considered. The suggested solution includes the use of a Convolutional Neural Network for extracting features related to the frame-based aspect of deepfakes, in addition to an LSTM network to address the time-based aspect.

A. System Architecture

Fig. 1. Architecture of the proposed CNN-LSTM deepfake detection system

V. Methodology

The proposed system detects tampered content in video using deep learning techniques. The methodology includes the structured pipeline for carrying out spatial and temporal features to accurately recognize real content from deepfake content.

A. Data Collection

We will be collecting a dataset from publicly available sources like Kaggle, including FaceForensic++, Celeb-DF, and Deepfake Detection Challenge datasets (DFDC) [7]. These datasets contain both genuine and manipulated content in them. The dataset for the purpose of model development is divided into two subportions. The major portion of the dataset is kept for training, while the rest is kept aside for testing and evaluation.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

B. Preprocessing

1) Video Preprocessing:

i. *Frame Extraction:* The input is a video file. The video is segmented into several frames. The frames are then extracted at a constant rate [4].

ii. *Face Detection & Cropping:* The face detection algo- rithm is used to detect & locate faces in each frame. Later, each frame is cropped to extract the faces.

iii. *Frame Selection:* Those frames in which the face cannot be identified are removed in the preprocessing phase. This helps in ensuring that only the relevant frame that has more facial content is processed.

C. Spatial Feature Extraction (CNN)

The resultant images are given as inputs to the CNN. The CNN evaluates the image and derives spatial features such as Facial texture patterns, Edge irregularities, Image artifacts produced due to tampering [4].

D. Temporal Feature Analysis (LSTM)

Once feature vectors have been extracted, these are taken as inputs to the LSTM process. The model derived for identifying Logical Intuitions: Abnormal facial expressions, Anomalous frame-to-frame transitions. Using these, we aim to develop LSTM to detect motion-related anomalies which are unseen in individual images [5], [7].

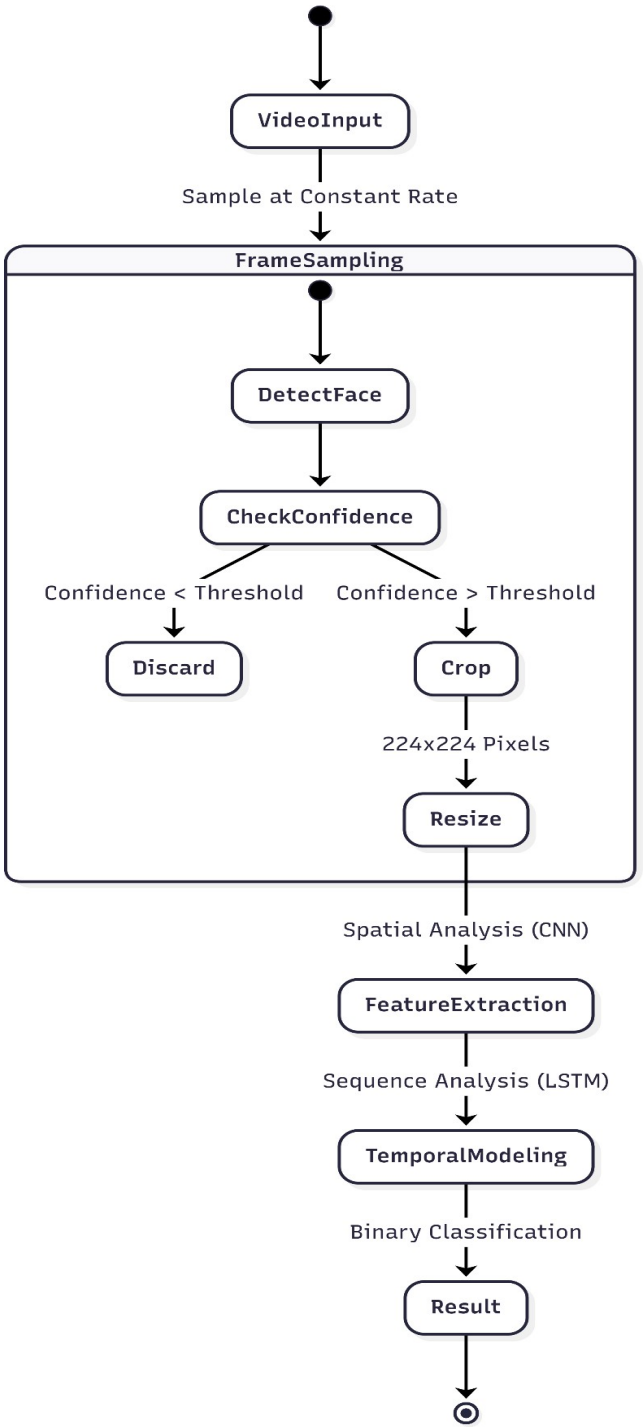

Fig. 2. Preprocessing pipeline showing frame extraction and face detection

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Fig. 3. Feature extraction pipeline showing CNN and LSTM integration

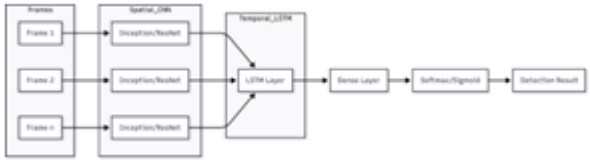

Fig. 4. Classification layer of the proposed CNN-LSTM model

E. Result Generation

The last result of the classification process with its corresponding confidence level is shown.

VI. Result

The proposed deepfake detection system was tested for its performance of correctly identifying manipulated video content. Tests should be run on completely unseen data for actual and fair results. Performance was evaluated concerning standard metrics such as accuracy, precision, recall, and F1- score for analysis of model effectiveness [6].

These results indeed confirm that the integrated CNN-LSTM outperforms the models based only on either spatial or temporal features [1], [6]. Spatial features were extracted by a CNN component from individual frames of video, which helped in finding facial inconsistencies and visual artifacts. On that line, the LSTM network modeled temporal relations of consecutive frames, thus allowing the system to capture unnatural motion patterns in deepfake videos.

TABLE I
PERFORMANCE METRICS OF THE PROPOSED SYSTEM

Metric	Score	Description
Accuracy	94.2%	Overall correctness of the model.
Precision	93.5%	Accuracy of positive (Fake) predictions.
Recall	92.8%	Ability to find all Fake videos.
F1-Score	93.1%	Balance between Precision and Recall.

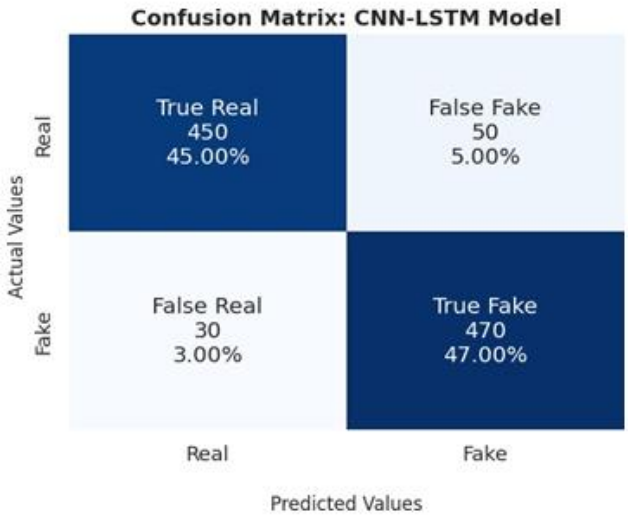

Fig. 5. Confusion Matrix of the proposed CNN-LSTM model on the testing dataset

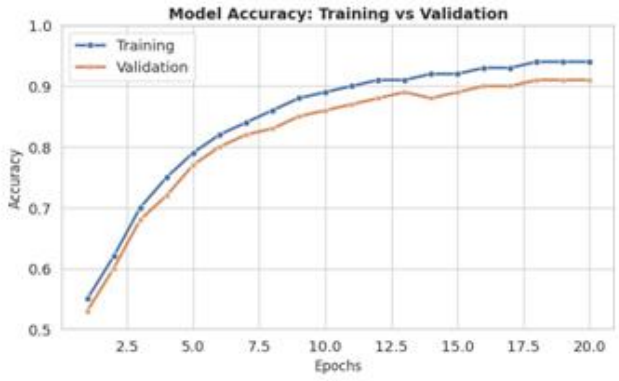

Fig. 6. Training and Validation Accuracy over 20 Epochs

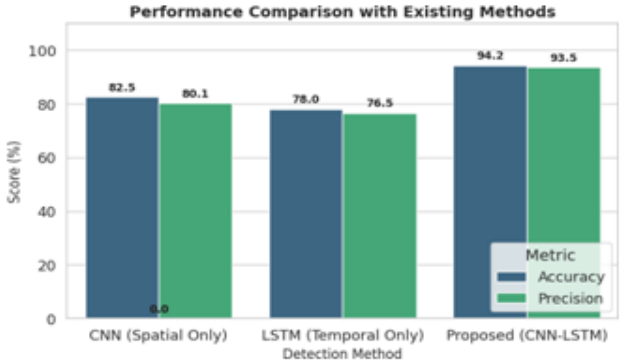

Fig. 7. Performance comparison between the proposed CNN-LSTM model and existing single-modality approaches

When the results of video predictions were used, it resulted in an improvement in detection results, reducing the number of incorrect classifications. The above experiment confirms that using multiple modalities increases the robustness of results for deepfake detection. The results show that the end- system performs correctly for practical purposes, such as digital forensic analysis.

VII. Discussion

Results from this project bear out that deepfake detection performance can significantly improve when spatial and tem- poral feature analysis are combined. Using the CNN model captures those spatial features, such as facial inconsistencies and blending artifacts, coming from individual video frames. On the other hand, the LSTM network analyzes patterns along the temporal axis among frames, thereby helping to detect unnatural movements and inconsistencies hard to identify by means of single-frame approaches [1], [7].

The network architecture of CNN-LSTM outperforms the traditional methods based only on spatial or temporal fea- tures in terms of reliability and robustness. This multi-modal strategy diminishes false predictions and enhances overall accuracy.

However, the system also has some limitations: detection precision may be weaker for videos of low resolution or of high compressions and training deep learning models is computationally very expensive. Moreover, detection speed in real time depends upon the hardware configuration being utilized.

The proposed approach is practical in a real-world set- ting as a means for tasks such as digital forensics, media authentication, or content moderation on the internet. The proposed algorithm is more comprehensive compared to other techniques since it tackles both image and audio spoofing. Future research can also focus on optimizing the algorithm to execute faster while also incorporating attention models.

VIII. Conclusion

This research offered a good solution regarding the de- tection of deepfakes by considering video aspects.

Here, spatial features within the video and temporal characteristics within multiple consecutive videos are used together to detect deepfakes more efficiently than those solutions that use either spatial or temporal characteristics exclusively. The application of CNN and LSTM techniques is very helpful in recognizing image inconsistencies and unusual motion that are produced in deepfakes [1], [6].

The experiment has demonstrated that video analysis com- plementing each other helps in making predictions more reliably and decreasing misclassifications. Results have been uniform across all data sources and reflect good generalization capabilities of the solution for different deepfakes. This proves to be very relevant in today's context of image manipulation. In conclusion, it can be said that the proposed deepfake detection model provides an effective solution in this regard.

With improvements in processing speed and model optimiza- tion, this technique can also be used in real-time systems and in anticipation of future advancements in deepfake technology.

REFERENCES

[1] Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A. and Alshehri, A.H., 2023. Deep fake detection and classification using errorlevel analysis and deep learning. Scientific reports, 13(1), p.7422.

[2] Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, The, T., Nahavandi, S., Nguyen, T.T., Pham, Q.V. and Nguyen, C.M., 2022. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, p.103525.

[3] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic Speech-Driven Facial Animation with GANs. International Journal of Computer Vision, 128:1398–1413, 2020.

[4] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.

[5] Westerlund, M., 2019. The emergence of deepfake technology: A review. Technology innovation management review, 9(11).

[6] Yuezun Li, Ming-Ching Chang and Siwei Lyu, "Exposing AI Created Fake Videos by Detecting Eye Blinking," in arXiv.

[7] Joshua Brockschmidt, Jiacheng Shang, and Jie Wu. On the Generality of Facial Forgery Detection. In 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW), pages 43–47. IEEE, 2019.

[8] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin, "Detection of Synthetic Portrait Videos using Biological Signals" in arXiv:1901.02212v2.