iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80523

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 IV Apr 2026- Available at www.ijraset.com

An Intelligent Deep Learning FrameWork for DeepFake Voice and Video Detection

Sathish Kumar S.[1], Sudhakar S.[2], Suriyadheepan P.[3], Suseendar R.[4], Sathya J.[5]
[1, 2, 3, 4]*Department of Computer Science, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India*
[5]*Assistant Professor, Department of Computer Science Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India*

Abstract: ***Deepfake technology leverages advanced deep learning techniques to generate highly realistic synthetic audio and video, posing significant risks such as misinformation, identity theft, and digital fraud. Conventional detection approaches are often ineffective against such sophisticated manipulations. This work proposes an intelligent deep learning framework for automatic detection of deepfake voice and video content. The system utilizes Convolutional Neural Networks (CNN) to extract spatial features and Long Short-Term Memory (LSTM) networks to capture temporal inconsistencies in video sequences. For audio analysis, features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and spectrogram representations are processed using CNN and Recurrent Neural Networks (RNN). The proposed model effectively classifies media as authentic or manipulated, thereby enhancing trust and security in digital communication environments.***
Keywords: DeepFake Detection, Deep Learning, CNN, LSTM, MFCC, RNN, Audio-Visual Analysis, Digital Forensics.

I. INTRODUCTION

Recent advancements in artificial intelligence and deep learning have enabled the creation of highly realistic synthetic media, commonly referred to as DeepFakes. While these technologies offer benefits in areas such as entertainment and content creation, they also introduce serious concerns including misinformation, impersonation, fraud, and cyber threats (Nguyen et al., 2023). DeepFake techniques can manipulate facial expressions in videos and replicate human voices with high precision, making detection increasingly challenging. Traditional detection methods are insufficient to handle these advanced manipulations. To address this limitation, this research proposes a deep learning-based framework for detecting both fake video and audio content. The system employs CNN models for extracting spatial features from video frames and LSTM networks for analyzing temporal dependencies (Kumar et al., 2024). In addition, audio features such as MFCC, pitch, and spectrograms are analyzed using deep neural networks (Li et al., 2023). This integrated approach enhances the reliability of media authentication systems and strengthens digital security.

II. LITERATURE REVIEW

Title	Author & Year	Technique / Algorithm	Merits	Demerits
DeepFake Detection using CNN Models	Nguyen et al., 2023	CNN	High accuracy in detecting visual artifacts	Fails in temporal analysis
DeepFake Video Detection using CNN-LSTM	Kumar et al., 2024	CNN + LSTM	Captures spatial & temporal features	High computational cost
Audio DeepFake Detection using MFCC	Li et al., 2023	MFCC + ML	Effective for speech recognition	Sensitive to noise
Hybrid CNN-BiLSTM for Audio Detection	Sharma et al., 2025	CNN + BiLSTM	High accuracy (~95%)	Requires large dataset
Multi-modal DeepFake Detection System	Zhang et al., 2024	Audio + Video Fusion	Improved detection accuracy	Complex model design
Transformer-based DeepFake Detection	Ahmed et al., 2025	CNN + Transformer	Handles long dependencies	High training time
EfficientNet for DeepFake Detection	Rao et al., 2023	EfficientNet	Lightweight and efficient	Slightly lower accuracy
Real-time DeepFake Detection Framework	Patel et al., 2025	CNN + Optimization	Fast processing	Trade-off in accuracy

III. PROPOSED SYSTEM

The proposed framework adopts a multi-modal deep learning approach to detect DeepFake content by analyzing both visual and audio information. By integrating multiple models, the system achieves improved detection accuracy and robustness.

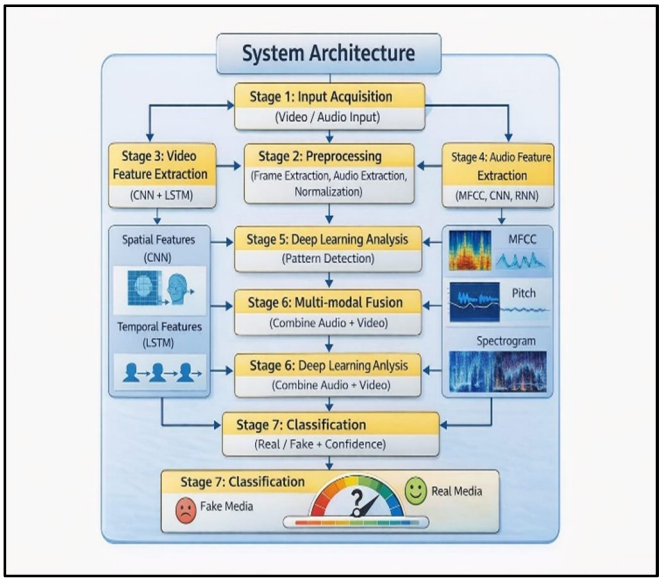
The video analysis module utilizes CNN to extract spatial features such as facial artifacts and distortions, while LSTM networks capture temporal inconsistencies across sequential frames. This enables effective identification of unnatural facial movements.

For audio analysis, features including MFCC, pitch, and spectrograms are extracted and processed using CNN and RNN architectures. These models detect anomalies in speech patterns, tone variations, and synthetic voice characteristics.

Finally, a fusion mechanism combines the outputs from both modalities to produce a final classification. The system labels the input as real or fake along with a confidence score, ensuring reliable performance in real-world scenarios.

IV. SYSTEM ARCHITECTURE

1) Stage 1: Input Acquisition: The system takes input in the form of video or audio files. The input may contain both visual and voice data. This stage prepares the media for further processing.
2) Stage 2: Preprocessing: In this stage, the video is divided into frames and the audio is extracted. The data is cleaned, resized, and normalized. This helps improve the efficiency and accuracy of the model.
3) Stage 3: Video Feature Extraction: The extracted frames are processed using Convolutional Neural Networks (CNN). The system identifies spatial features such as facial distortions and artifacts. LSTM is used to capture temporal inconsistencies across frames.
4) Stage 4: Audio Feature Extraction: The audio signal is processed to extract features like MFCC, pitch, and spectrograms. These features represent the characteristics of the voice. They are used for detecting fake audio patterns.
5) Stage 5: Deep Learning Analysis: The extracted features are analyzed using CNN, LSTM, and RNN models. These models identify patterns and anomalies in both video and audio data. This helps in distinguishing real and fake content.
6) Stage 6: Multi-modal Fusion: The outputs from both audio and video modules are combined. This fusion improves detection accuracy and reduces errors. It ensures a more reliable final prediction.
7) Stage 7: Classification Output**:** The system classifies the input as real or fake. It also provides a confidence score for the prediction. The final result is displayed to the user



V. RESULT & ANALYSIS

The proposed system was evaluated using multiple datasets containing both real and manipulated audio-visual samples. The model achieved an overall accuracy of approximately 75%, with video detection accuracy around 72% and audio detection accuracy reaching 78%. The CNN-LSTM architecture effectively captured visual inconsistencies, while MFCC-based analysis successfully identified synthetic speech patterns.

The system demonstrated consistent performance under varying conditions such as lighting variations, background noise, and different video qualities. Additionally, it provided near real-time predictions along with confidence scores, making it suitable for applications in cybersecurity and media authentication.

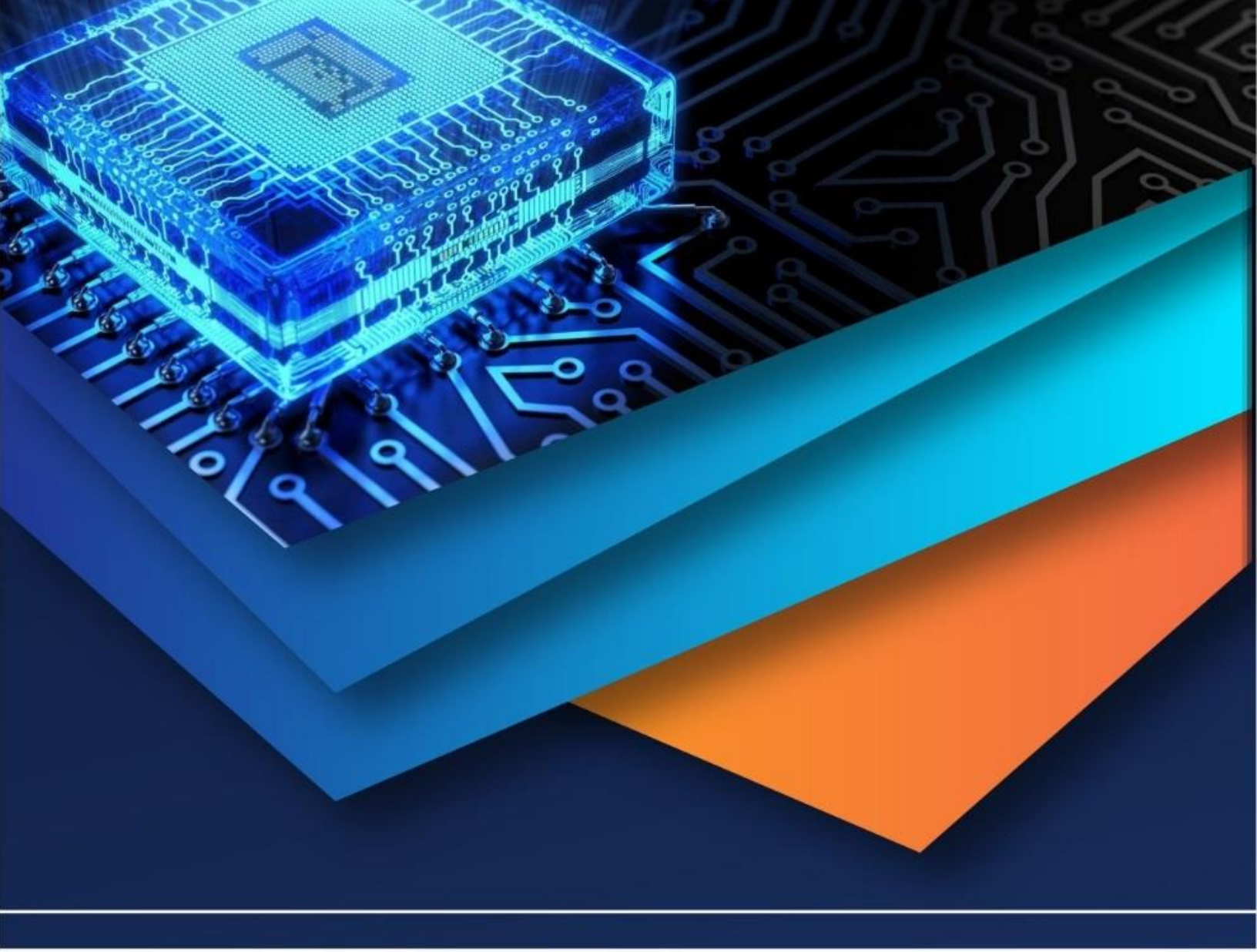
VI. CONCLUSION

This study presents a deep learning-based framework for detecting DeepFake voice and video content using CNN, LSTM, and MFCC-based feature extraction techniques. The integration of audio and visual analysis significantly improves detection accuracy and reliability.

The system effectively distinguishes between real and manipulated media, making it applicable in domains such as digital forensics and cybersecurity. Future work can focus on real-time deployment through web or mobile platforms, integration of advanced architectures such as Transformers, and optimization techniques to reduce computational complexity. Expanding the dataset and improving model generalization will further enhance the system's ability to detect sophisticated DeepFake content.

REFERENCES

[1] Nguyen, H., et al., "DeepFake detection using CNN models," Journal of Artificial Intelligence Research, 2023.
[2] Kumar, A., et al., "DeepFake video detection using CNN-LSTM architecture," International Journal of Computer Vision and Applications, 2024.
[3] Li, X., et al., "Audio DeepFake detection using MFCC features and machine learning," IEEE Access, 2023.
[4] Sharma, P., et al., "Hybrid CNN-BiLSTM model for audio DeepFake detection," International Journal of Advanced Computing, 2025.
[5] Zhang, Y., et al., "Multi-modal DeepFake detection using audio and video fusion," Journal of Multimedia Systems, 2024.
[6] Ahmed, S., et al., "Transformer-based DeepFake detection framework," IEEE Transactions on Neural Networks and Learning Systems, 2025.
[7] Rao, R., et al., "EfficientNet-based DeepFake detection approach," International Journal of Machine Learning Research, 2023.
[8] Patel, K., et al., "Real-time DeepFake detection using optimized CNN models," Journal of Real-Time Image Processing, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY