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.2026.80240

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], Suriyadheepan P[2], Suseendar R[3], Sudhakar[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 uses advanced deep learning models to create realistic fake audio and video, leading to threats like fraud, misinformation, and identity theft. Traditional detection methods fail to identify such high-quality manipulations. This project presents a deep learning-based system to detect deepfake voice and video content automatically. CNN and LSTM models are used to identify spatial and temporal inconsistencies in videos. For voice detection, features like MFCC, pitch, and spectrograms are analyzed using CNN and RNN. The system accurately classifies media as real or fake, enhancing digital trust and security.***

Keywords: DeepFake Detection, Deep Learning, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Mel-Frequency Cepstral Coefficients (MFCC), Recurrent Neural Network (RNN), Audio-Visual Analysis, Digital Forensics, Media Authentication.

I. INTRODUCTION

In recent years, the rapid advancement of artificial intelligence and deep learning technologies has led to the emergence of highly realistic synthetic media known as DeepFakes. While such technologies have useful applications in entertainment and media, they also pose serious threats including misinformation, identity theft, fraud, and cybercrime. DeepFake videos can manipulate facial expressions and identities, while voice cloning techniques can replicate a person's speech with high accuracy, making it difficult to distinguish between real and fake content. Traditional detection methods are no longer effective against these sophisticated manipulations. To overcome these challenges, this project proposes an intelligent deep learning-based framework for detecting DeepFake voice and video content. The system uses Convolutional Neural Networks (CNN) to extract spatial features from video frames and Long Short-Term Memory (LSTM) networks to analyze temporal inconsistencies. For audio detection, features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and spectrograms are analyzed using deep learning models. The proposed system classifies media content as real or fake, contributing to enhanced digital security and trust in modern communication systems.

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 system is a deep learning-based framework designed to detect DeepFake voice and video content. It uses a multi-modal approach to analyze both visual and audio data.

The system improves accuracy by combining different models. It provides a reliable solution for identifying fake media.

The video detection module uses Convolutional Neural Networks (CNN) to analyze frames and detect visual inconsistencies. Long Short-Term Memory (LSTM) models are used to capture temporal changes between frames. This helps in identifying unnatural movements and facial distortions.

The audio detection module extracts features such as MFCC, pitch, and spectrograms from the input. These features are analyzed using CNN and RNN models to detect fake voice patterns. The system identifies unnatural tone and speech inconsistencies.

Finally, a multi-modal fusion approach combines audio and video results to give the final output.

The system classifies the media as real or fake with a confidence score. This improves accuracy and ensures better performance in real-world applications.

IV. SYSTEM ARCHITECTURE

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.

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.

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.

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.

Stage 5: Deep Learning Analysis

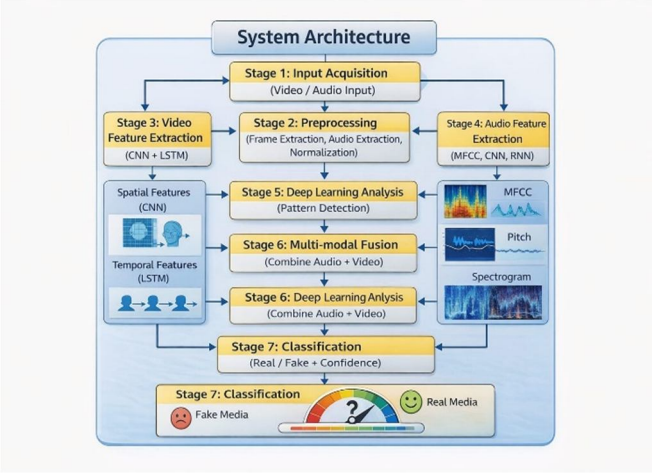
The extracted features are analyzed using CNN, LSTM, and RNN models. These models identify patterns and anomalies in both video and audio data. This helps in distinguishing real and fake content.

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.

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



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

V. RESULT & ANALYSIS

The proposed Deep Fake detection system was evaluated using various real and manipulated audio-video datasets. The system achieved an overall accuracy of approximately 75%, with the video detection module reaching around 72% accuracy and the audio detection module achieving about 78% accuracy. The CNN and LSTM models effectively detected visual inconsistencies, while MFCC-based audio analysis successfully identified synthetic voice patterns. The system also performed reliably under different conditions such as varying lighting, background noise, and video quality. Additionally, it provided near real-time results along with a confidence score, making it suitable for practical applications in cybersecurity and media verification.

VI. CONCLUSION

The proposed deep learning-based system effectively detects DeepFake voice and video content by using CNN, LSTM, and MFCC-based feature extraction.

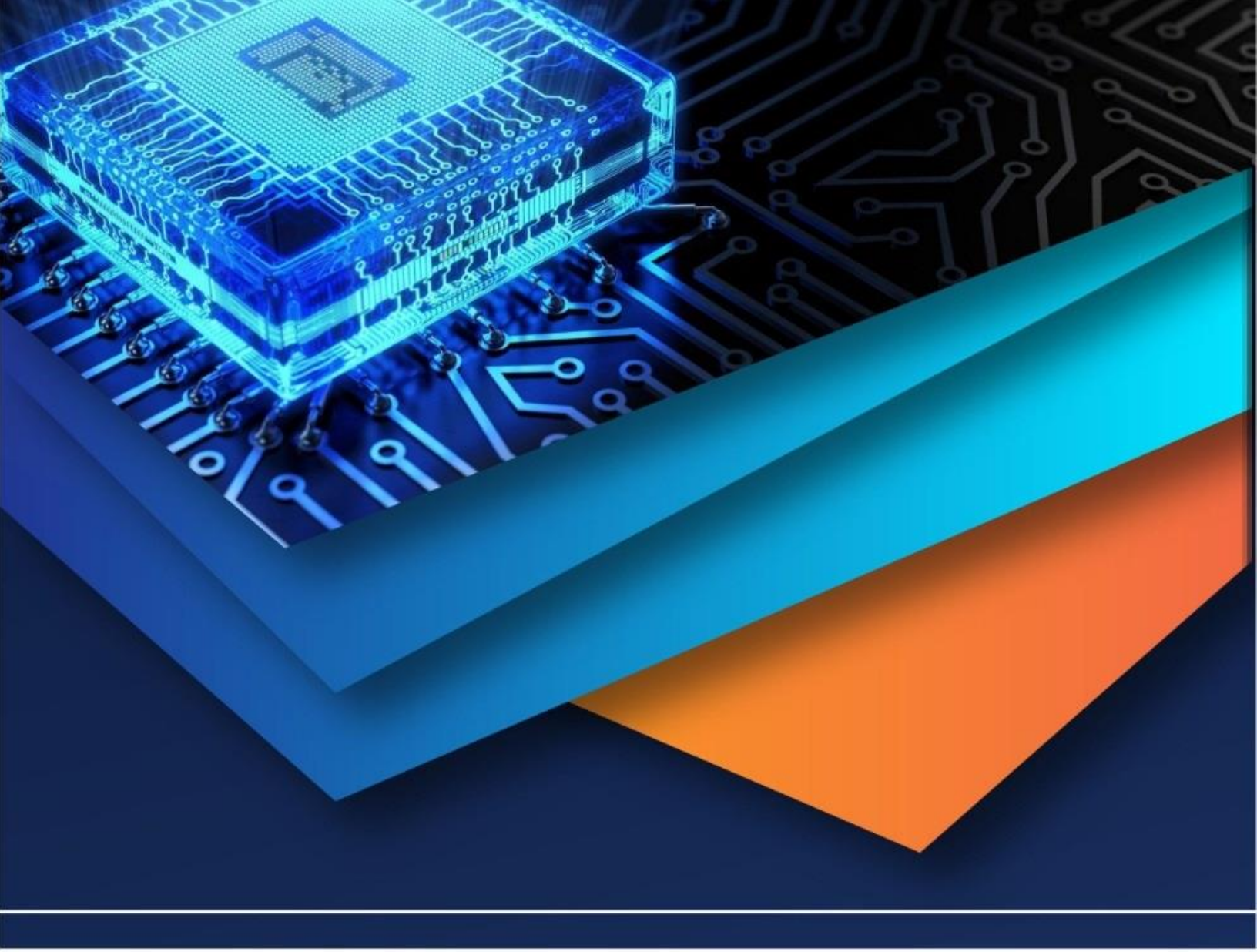
The multi-modal approach improves accuracy and reliability by combining both audio and video analysis. The system successfully classifies media as real or fake and performs well under different conditions, making it suitable for applications in cybersecurity and digital forensics.

In the future, the system can be enhanced for real-time detection and deployed as a web or mobile application. Further improvements can include the use of advanced models such as Transformers and optimization techniques to reduce computational cost.

Expanding the dataset and improving generalization will also help in detecting more 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