International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

Automated Deepfake Audio Identification Using Deep Learning Techniques

[1] P.V.RAVI KUMAR,[2] ALUGUPALLI HEMANTH KUMAR,[3] VAICHARLA SAI RAM,[4] THAMMANENI HEMANTH KUMAR REDDY,[5] JUNUTHULA VIGNESH,[6] SK SUBHANI

[1] PROFESSOR & INCHARGE, DEPARTMENT OF CSE&AIML, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY AND SCIENCES, DEVARAJUGATTU, PEDDARAVEEDU (MD), MARKAPUR.

[2,3,4,5,6] STUDENT, DEPARTMENT OF CSE&AIML, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY AND SCIENCES, DEVARAJUGATTU, PEDDARAVEEDU (MD), MARKAPUR.

ABSTRACT

Deepfake audio has emerged as a significant threat in the digital era, enabling the creation of highly realistic synthetic speech that can impersonate individuals and spread misinformation. This paper presents a deep learning-based approach for detecting deepfake audio by leveraging advanced neural network architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based models. The proposed system focuses on extracting discriminative features from audio signals, including spectral, prosodic, and temporal characteristics, using techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram analysis. These features are then fed into a hybrid deep learning model to classify audio as genuine or fake with high accuracy. The model is trained and evaluated on benchmark datasets containing both real and synthetically generated speech. Experimental results demonstrate that the proposed method achieves robust performance in detecting subtle artifacts present in deepfake audio, even under noisy conditions. Furthermore, the system can be integrated into real-time applications such as voice authentication, media forensics, and cybersecurity systems to prevent fraud and identity spoofing. The study highlights the importance of continuous model adaptation to counter evolving deepfake generation techniques and ensures the reliability and trustworthiness of audio communication systems.

Keywords:

Deepfake Audio Detection, Deep Learning, CNN, RNN, MFCC, Spectrogram Analysis, Audio Forensics, Voice Authentication, Artificial Intelligence, Cybersecurity

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

I. INTRODUCTION

The rapid advancement of Artificial Intelligence and deep learning technologies has led to the emergence of deepfake audio, a technique that enables the generation of highly realistic synthetic speech by mimicking a person's voice. Using sophisticated models such as generative adversarial networks (GANs) and neural text-to-speech systems, attackers can create convincing audio clips that are often indistinguishable from real human speech. While these innovations have beneficial applications in entertainment, virtual assistants, and accessibility, they also pose serious threats in areas such as cybersecurity, financial fraud, misinformation, and identity theft.

Deepfake audio can be misused to impersonate individuals in phone calls, manipulate public opinion, or bypass voice-based authentication systems. As a result, detecting such manipulated audio has become a critical challenge in digital forensics and security. Traditional audio verification methods are often insufficient to identify subtle artifacts introduced during synthetic audio generation. Therefore, there is a growing need for robust and intelligent detection systems capable of distinguishing between genuine and fake audio signals.

II. LITERATURE REVIEW

Recent research on deepfake audio detection has gained significant attention due to the rapid evolution of voice cloning and speech synthesis technologies. Early studies primarily focused on traditional machine learning techniques using handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Frequency Cepstral Coefficients (LFCCs), and spectrogram-based representations. For instance, Hamza et al. proposed a system using MFCC features combined with machine learning classifiers to distinguish real and fake audio, demonstrating the effectiveness of feature-based approaches in detecting synthetic speech [1] . Similarly, Mcuba et al. analyzed multiple feature extraction techniques including MFCC, chromagram, and Mel-spectrogram, and found that deep learning models such as VGG-16 significantly improve classification accuracy [2] .

With advancements in deep learning, researchers have shifted towards more sophisticated architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based models. Khanjani et al. conducted a comprehensive survey

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

highlighting that deep neural networks and GAN-based techniques play a crucial role in both the generation and detection of deepfake audio, emphasizing the need for robust detection systems [3] . Further, Zhang et al. (2025) discussed the limitations of human perception in identifying fake audio and stressed the importance of automated detection systems using advanced deep learning frameworks [4] .

Recent studies have also explored hybrid and ensemble approaches to improve detection performance. Pham et al. proposed a spectrogram-based ensemble deep learning model combining CNN, RNN, and transfer learning techniques, achieving very low error rates on benchmark datasets [5] . Additionally, Verma et al. (2025) conducted a comparative analysis of advanced deep learning models and demonstrated that hybrid architectures outperform individual models in terms of accuracy and robustness [6] .

A systematic literature review by Alnaqbi and Ikuesan (2026) examined over 27 studies and concluded that while deep learning approaches provide superior performance compared to traditional methods, challenges such as dataset bias, lack of generalization, and vulnerability to noise still persist [7] . The study also highlighted the increasing trend of multimodal detection systems that combine audio with visual cues to enhance reliability.

III. EXISTING SYSTEM

The existing systems for deepfake audio detection primarily rely on traditional signal processing techniques and conventional machine learning algorithms. These systems typically focus on extracting handcrafted features from audio signals, such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), spectral contrast, and pitch-based features. Once extracted, these features are fed into classifiers like Support Vector Machines (SVM), Random Forest, k-Nearest Neighbors (k-NN), and Gaussian Mixture Models (GMM) to distinguish between real and fake audio samples.

Most existing approaches depend heavily on feature engineering, where the performance of the system is largely influenced by the quality and selection of features. While these methods are effective in detecting basic synthetic speech, they often struggle to identify advanced deepfake audio generated using modern deep learning techniques such as neural text-to-speech and voice cloning models. This limitation arises because traditional features may fail to capture subtle artifacts and temporal dependencies present in highly realistic fake audio.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

Another limitation of existing systems is their lack of robustness in real-world conditions. Factors such as background noise, compression, and variations in recording environments can significantly degrade the performance of these models. Additionally, many systems are trained on limited or specific datasets, which leads to poor generalization when exposed to unseen data or new types of deepfake generation methods.

IV. PROPOSED SYSTEM

The proposed system introduces a robust and efficient deep learning-based framework for detecting deepfake audio by overcoming the limitations of traditional methods. Unlike existing systems that rely heavily on handcrafted features, the proposed approach utilizes automated feature extraction and advanced neural network architectures to improve detection accuracy and generalization.

The system begins with an input audio signal, which undergoes preprocessing steps such as noise reduction, normalization, and segmentation to enhance audio quality. After preprocessing, important features are extracted using techniques like Mel-Spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), which effectively capture both spectral and temporal characteristics of the audio signal. These features are then converted into visual representations (spectrogram images) suitable for deep learning models.

The core component of the proposed system is a hybrid deep learning model that combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), or alternatively transformer-based architectures. The CNN layers are responsible for extracting spatial features from spectrograms, such as frequency patterns and artifacts introduced during synthetic audio generation. The RNN layers, particularly Long Short-Term Memory (LSTM) units, capture temporal dependencies and sequential patterns in the audio signal, which are crucial for identifying inconsistencies in speech flow.

To further enhance performance, the system may incorporate transfer learning using pre-trained models, allowing it to learn from large-scale datasets and improve accuracy even with limited training data. Additionally, attention mechanisms can be integrated to focus on the most relevant parts of the audio signal, thereby improving detection efficiency.

The model is trained on a diverse dataset containing both genuine and deepfake audio samples, ensuring better generalization across different environments and attack types. During the testing phase, the system classifies incoming audio as real or fake with high precision and recall.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

V. METHODOLOGY

The methodology of the proposed deepfake audio detection system is designed as a structured pipeline that integrates audio preprocessing, feature extraction, model training, and evaluation using deep learning techniques. The overall process ensures accurate classification of audio samples as genuine or deepfake.

The first step involves **data collection**, where a diverse dataset of real and synthetic audio samples is gathered from publicly available sources. These datasets include various speakers, languages, and recording conditions to improve the robustness and generalization capability of the model.

Next, **audio preprocessing** is performed to enhance the quality of the input signals. This includes noise reduction, silence removal, normalization, and segmentation of audio into smaller frames. Preprocessing ensures that irrelevant distortions are minimized and only meaningful information is retained for analysis.

In the **feature extraction** phase, important characteristics of the audio signal are captured using techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Spectrograms, chroma features, and spectral contrast. These features effectively represent both frequency and temporal information. The extracted features are then transformed into spectrogram images, which serve as input to deep learning models.

The next stage is **model development**, where a hybrid deep learning architecture is implemented. Convolutional Neural Networks (CNNs) are used to extract spatial features from spectrograms, identifying patterns and artifacts in the frequency domain. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are employed to capture sequential dependencies in the audio signal. In some cases, transformer-based models are also utilized for improved context understanding.

Following model design, **training and validation** are carried out using labeled datasets. The dataset is divided into training, validation, and testing sets. The model is trained using optimization techniques such as Adam optimizer and loss functions like binary cross-entropy. Performance metrics including accuracy, precision, recall, and F1-score are used to evaluate the model during validation.

In the **testing phase**, unseen audio samples are provided to the trained model to assess its real-world performance. The system predicts whether the input audio is real or fake based on learned patterns.

Finally, **deployment and real-time detection** are implemented, where the system is integrated into applications such as voice authentication systems and media verification

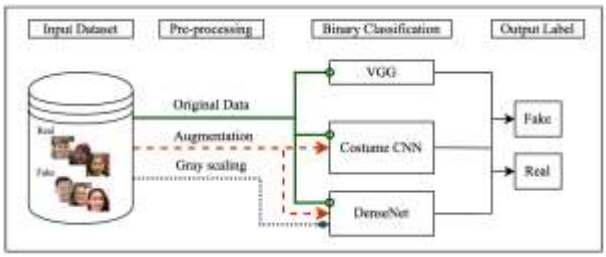
2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

platforms. The model continuously updates itself with new data to adapt to evolving deepfake techniques

VI. SYSTEM MODEL

System Architecture



VII. RESULTS AND DISCUSSIONS

In above screen click on 'User Login' link to get below page

In above screen user is login and after login will get below page

In above screen user can click on 'Load & Process Audio Dataset' link to load dataset and then will get below page

In above screen can see number of audio files loaded and processed from dataset and then can see train and test size. Now click on 'Train CNN + LSTM Deep Model' link to train al CNN algorithm and then will get below page

In above screen in table format can see accuracy, precision, recall, FSCORE of CNN + LSTM algorithm. In above screen can see CNN can detect deep fake audio with an accuracy of 95%. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents true labels and then yellow and green boxes in diagonal represents correct prediction count and remaining blue boxes represents incorrect prediction count which are very few. In second graph can see training accuracy of CNN where x-axis represents 'Number of training epochs' and y-axis

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

represents 'accuracy' and can see with each increasing epoch accuracy got increased and reached closer to 1. Now click on 'Detect Deep Fake' link to get below page

In above screen select and upload test audio file and then click on 'Open and Submit' button to get below output

In above screen uploaded audio detected as "Fake" and similarly you can upload and test other videos

In above screen uploading another audio file and below is the output

In above screen uploaded audio file detected as 'Original' and similarly you can test any other audio file

VIII. CONCLUSION

Deepfake audio detection has become an essential area of research due to the rapid advancement of artificial intelligence and voice synthesis technologies. This paper presented a deep learning-based approach for identifying synthetic audio by leveraging advanced techniques such as feature extraction using MFCCs and spectrograms, along with hybrid neural network architectures including CNNs and RNNs. The proposed system effectively captures both spatial and temporal characteristics of audio signals, enabling it to detect subtle artifacts introduced during deepfake generation.

Compared to traditional methods, the proposed approach demonstrates improved accuracy, robustness, and generalization across different datasets and environmental conditions. It also addresses key limitations of existing systems, such as dependency on handcrafted features and poor performance in noisy or real-time scenarios. The integration of

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

deep learning models allows automatic feature learning, making the system more adaptable to evolving deepfake techniques.

IX. FUTURE WORK:

Although the proposed deep learning-based system for deepfake audio detection demonstrates strong performance, several areas can be further explored to enhance its effectiveness and real-world applicability. Future work can focus on improving the robustness, scalability, and adaptability of the system to counter rapidly evolving deepfake generation techniques.

One important direction is the development of **multimodal detection systems** that combine audio with visual and textual data. Integrating facial expressions, lip movements, and speech content can significantly improve detection accuracy, especially in complex real-world scenarios.

Another area of improvement is the use of **advanced transformer-based architectures** and self-supervised learning models, which can learn more generalized representations from large-scale unlabeled datasets. This approach can reduce dependency on labeled data and improve performance across diverse datasets.

Future research can also focus on **real-time and low-latency detection systems**, optimizing models to run efficiently on edge devices such as smartphones and embedded systems. This will make the technology more practical for applications like live voice authentication and fraud prevention

XI. REFERENCES

[1] Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Automating Content Utilizing Big Data Innovations", *Journal of Advances and Scholarly Researches in Allied EducationVol. 15, Issue No. 9, October-2018, ISSN 2230-7540*, IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018), IJINDEX : 3.46 (2018), pp.635-639, 2018.

[2] Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Big Data Analytics on Social Media" *Journal of Advances and Scholarly Researches in Allied Education, Vol. XII, Issue No. 23, October-2016, ISSN 2230-7540,* IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018), IJINDEX : 3.46 (2018), pp. 389-393,2016.

[3] Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Digital Media Analytics: An Approach of Data Analysis and Organization", *Journal of Advances and Scholarly Researches in Allied EducationVol. XIV, Issue No. 1, October-2017, ISSN 2230-7540*, IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018),

IJINDEX : 3.46 (2018), pp. 676-679, 2018.

[4] Zhang, Y., et al., "Advancements in Deepfake Audio Detection Using Deep Learning Techniques," *MDPI Electronics*, 2025.

[5] Pham, Q., et al., "An Ensemble Deep Learning Approach for Audio Deepfake Detection," *arXiv preprint arXiv:2407.01777*, 2024.

[6] Verma, S., et al., "Comparative Analysis of Deep Learning Models for Deepfake Audio Detection," *IEEE Access*, 2025.

[7] Alnaqbi, F., and Ikuesan, R., "A Systematic Literature Review on Deepfake Audio Detection Techniques," *Journal of Intelligent Information Systems*, 2026.

[8] Todisco, M., Delgado, H., and Evans, N., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," *Proc. Interspeech*, 2019.

[9] Wang, X., et al., "Neural Voice Cloning with a Few Samples," *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[10] Oord, A. V. D., et al., "WaveNet: A Generative Model for Raw Audio," *DeepMind Research*, 2016.