IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15377

Deepfake Audio Detection Using Hybrid Random Forest and Convolutional Neural Network Architecture

K Dharma Ratnam[1], P Kanaka Tulasi[2]

M. Tech Scholar, CSE, Pragati Engineering college, Kakinada, India[1]

Assistant Professor, CSE, Pragati Engineering college, Kakinada, India[2]

Abstract: The rapid evolution of speech synthesis and voice conversion technologies has enabled the generation of highly realistic synthetic speech, commonly referred to as deepfake audio. While such technologies offer innovative applications in media and accessibility, they also introduce serious threats to security, privacy, and information authenticity. This paper presents a hybrid deepfake audio detection system that combines classical machine learning and deep learning techniques to identify spoofed speech. The proposed framework integrates a Random Forest classifier trained on Mel-Frequency Cepstral Coefficients (MFCCs) and a Convolutional Neural Network (CNN) trained on Log-Mel Spectrogram representations. The system is implemented as a standalone desktop application with real-time visualization support. Experimental results demonstrate that the hybrid approach achieves high classification accuracy while maintaining computational efficiency suitable for consumer-grade hardware. The proposed solution aims to provide an accessible and reliable tool for combating synthetic audio misuse.

Keywords: Deepfake Audio, Audio Spoofing Detection, MFCC, CNN, Random Forest, Spectrogram Analysis, Anti-Spoofing

I. INTRODUCTION

Recent progress in deep neural networks has substantially enhanced the quality of synthetic speech generation systems. Modern text-to-speech and voice conversion frameworks can produce speech signals that closely resemble natural human voices, making it increasingly difficult to distinguish between authentic and artificially generated audio. Architectures such as generative adversarial networks and autoregressive models have enabled near-human-quality speech generation [1], [8]. While these technologies enhance human-computer interaction and digital media production, they also enable malicious impersonation attacks.

Deepfake audio refers to artificially generated speech that imitates the vocal characteristics of a specific individual. Such audio is typically produced using advanced text-to-speech (TTS) or voice conversion (VC) techniques that replicate pitch, tone, and speaking style of a target speaker. Modern neural vocoders such as WaveNet [8] and StarGAN-VC [9] produce speech that is perceptually indistinguishable from real human speech. This creates vulnerabilities in voice-based authentication systems and increases risks in social engineering, fraud, and misinformation dissemination [1]-[3].

The Automatic Speaker Verification Spoofing (ASVspoof) challenges [1]-[3] have demonstrated the growing complexity of synthetic speech detection. Traditional feature-based systems show reduced performance against modern neural synthesis techniques [4]. Consequently, research has shifted toward deep representation learning [5], [6].

This work proposes a hybrid detection architecture that combines:

1) A lightweight Random Forest (RF) classifier using MFCC features.
2) A Convolutional Neural Network (CNN) operating on Log-Mel spectrograms.

The key contributions of this paper are:

1) Development of a dual-model detection framework balancing speed and accuracy.
2) Implementation of a user-friendly desktop system for practical deployment.
3) Comparative evaluation of classical and deep learning approaches.
4) Integration of visual analytics for waveform and spectrogram inspection.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15377

II. RELATED WORK

Audio spoofing detection research has evolved through multiple phases.

A. Feature-Based Approaches

Early systems relied on handcrafted acoustic features such as MFCC and CQCC [4]. MFCCs, originally proposed for speech recognition [10], capture spectral envelope characteristics and remain widely used in anti-spoofing systems. Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) were commonly applied classifiers [2]. While computationally efficient, these approaches struggle with advanced neural vocoders.

B. Deep Learning Approaches

CNN-based models treat spectrograms as two-dimensional images [7]. By learning spatial patterns in time-frequency representations, CNNs effectively detect artifacts introduced during synthesis. Light CNN (LCNN) models reduce parameter count while maintaining performance [6]. Raw waveform models such as RawNet directly process one-dimensional signals [5], capturing phase information lost in magnitude spectrograms. Residual networks (ResNet) [7] have also been adapted for spoof detection tasks, improving generalization capability.

Despite high performance, deep models often require substantial computational resources. Therefore, combining classical and deep models offers a practical compromise.

III. PROPOSED SYSTEM ARCHITECTURE

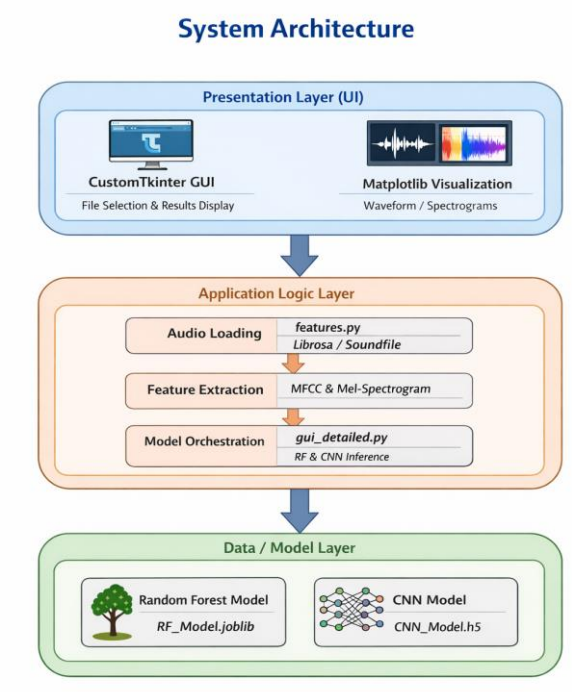

Fig. 1. Overall System Architecture

The overall architecture of the proposed deepfake audio detection system is illustrated in Fig. 1. The system is designed as a modular pipeline that processes audio inputs and performs classification using both classical machine learning and deep learning models. The architecture is composed of four primary components: an audio acquisition module, a feature extraction module, a classification module, and a visualization interface module.

The audio acquisition module accepts multiple audio formats including WAV, MP3, FLAC, and OGG files. Once the audio is loaded, preprocessing operations such as normalization and resampling are performed to ensure consistent signal quality. The processed audio signal is then passed to the feature extraction module where acoustic representations suitable for machine learning are generated.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15377

The feature extraction stage produces two types of representations: Mel-Frequency Cepstral Coefficients (MFCC) and Log-Mel spectrogram features. These representations capture spectral and temporal properties of speech signals that are useful for distinguishing between genuine and synthesized speech. The extracted features are subsequently fed into the classification module, where predictions are generated using either the Random Forest classifier, the Convolutional Neural Network model, or both simultaneously. Finally, the results are presented through the visualization interface, which displays prediction probabilities along with waveform and spectrogram visualizations for interpretability.

The complete processing flow is shown in Fig. 2.

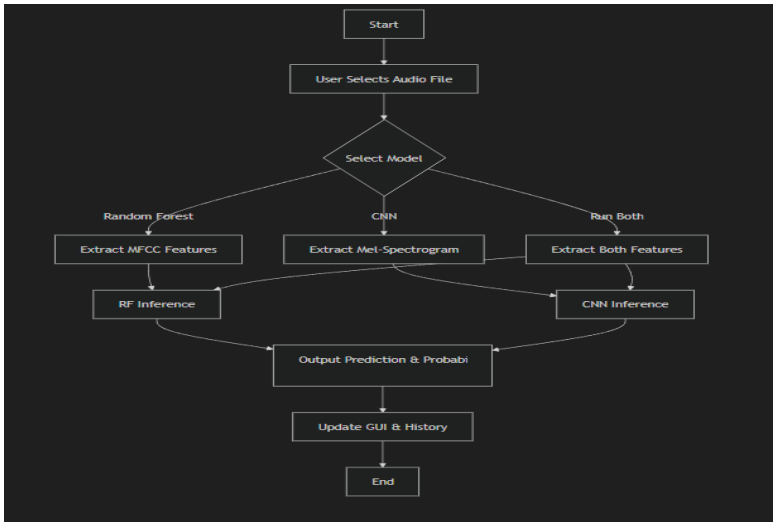

Fig. 2. System Processing Flowchart

The processing steps are as follows:

1) Load audio file (.wav, .mp3, .flac, .ogg)
2) Normalize and resample audio
3) Extract MFCC and Log-Mel features
4) Perform classification (RF / CNN / Both)
5) Display probability score and visualization

IV. FEATURE EXTRACTION

Feature extraction plays a crucial role in identifying the acoustic differences between natural human speech and artificially generated audio. In the proposed system, two complementary feature representations are employed: MFCC features for traditional machine learning models and Log-Mel spectrograms for deep learning-based analysis.

Mel-Frequency Cepstral Coefficients (MFCC) are among the most widely used acoustic features in speech processing because they approximate the frequency perception characteristics of the human auditory system [10]. The extraction process begins with a pre-emphasis stage that enhances high-frequency components of the signal. The audio waveform is then divided into short overlapping frames of approximately 25 milliseconds, followed by the application of a Hamming window to reduce spectral leakage. Each frame is transformed into the frequency domain using the Fast Fourier Transform (FFT), after which the power spectrum is mapped onto a Mel-scale filter bank to approximate human auditory perception. Logarithmic compression is applied to stabilize the dynamic range of the signal, and the Discrete Cosine Transform (DCT) is finally used to obtain decorrelated MFCC coefficients.

In the proposed implementation, forty MFCC coefficients are extracted from each audio sample, and global mean aggregation is applied to obtain a fixed-length feature vector of dimension forty. In addition to MFCC features, Log-Mel spectrograms illustrated in Fig. 3 are generated to capture detailed time–frequency patterns of the speech signal. These spectrograms are computed using the Short-Time Fourier Transform and mapped onto 128 Mel frequency bands. Logarithmic scaling and normalization are applied to standardize the input before it is provided to the CNN model. Previous studies have shown that synthetic speech often exhibits abnormal spectral smoothing and high-frequency artifacts, which can be effectively detected through spectrogram analysis [4].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15377

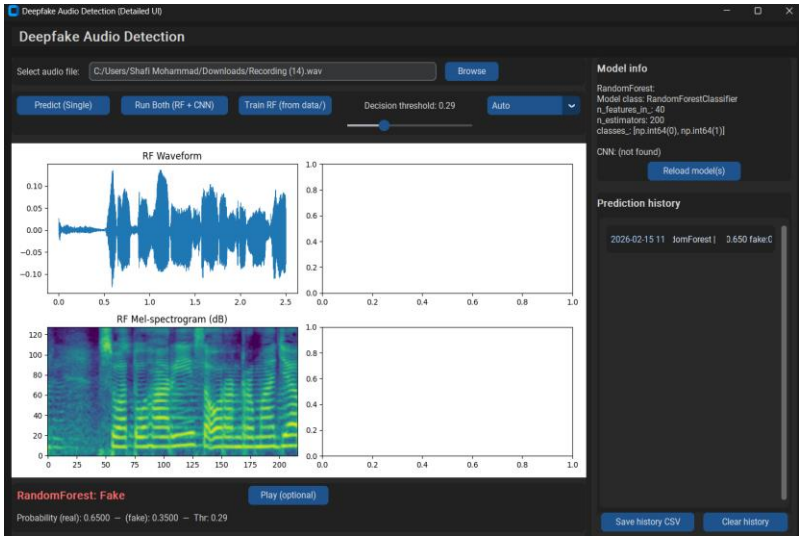

Fig. 3. Example Log-Mel Spectrogram Representation

V. CLASSIFICATION MODELS

The classification stage consists of two complementary models: a Random Forest classifier and a Convolutional Neural Network. The objective of combining these models is to balance computational efficiency with predictive accuracy.

The Random Forest classifier is a supervised ensemble learning method that constructs multiple decision trees during training and aggregates their predictions through majority voting. This model is trained using MFCC feature vectors extracted from the speech signals. In the proposed system, the Random Forest model uses 200 decision trees and employs the Gini impurity criterion to measure node splitting quality. Balanced class weights are applied to address potential class imbalance between real and synthetic speech samples. Due to its relatively low computational complexity, the Random Forest model provides fast inference and is suitable for lightweight deployment scenarios.

In contrast, the Convolutional Neural Network model operates on Log-Mel spectrogram representations of the audio signal. CNN architectures are particularly effective for analyzing spectrograms because they can capture spatial correlations in time–frequency patterns [7]. The network architecture used in this study consists of three convolutional layers with increasing numbers of filters, followed by a global average pooling layer and a fully connected dense layer. Batch normalization and dropout regularization are applied to improve generalization and reduce overfitting. The network is trained using the binary cross-entropy loss function and optimized with the Adam optimizer. The detailed architecture of the CNN model is presented in Table I.

TABLE I: CNN ARCHITECTURE

Layer	Filters	Kernel	Activation
Conv2D	32	3×3	ReLU
Conv2D	64	3×3	ReLU
Conv2D	128	3×3	ReLU
Global Avg Pool	-	-	-
Dense	128	-	ReLU
Output	1	-	Sigmoid

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15377

VI. EXPERIMENTAL SETUP

A. Dataset

The dataset used in this study contains two categories of audio samples:

1) Genuine human speech recordings
2) Synthetic speech generated using TTS and voice conversion models

The dataset was randomly divided into training and testing subsets using an 80:20 ratio. The CNN model was trained for 30 epochs with a batch size of 32. Early stopping was applied to prevent overfitting.

B. Hardware Configuration

TABLE II: HARDWARE CONFIGURATION

Component	Specification
CPU	Intel i5 / Ryzen 5
RAM	8–16 GB
GPU	Optional NVIDIA GTX 1060
OS	Windows 10/11

C. Evaluation Metrics

The performance of the proposed models was evaluated using several standard classification metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics provide a comprehensive evaluation of the models' ability to correctly distinguish between real and synthetic speech samples.

VII. RESULTS AND ANALYSIS

The experimental evaluation demonstrates that the proposed hybrid framework is effective in detecting deepfake audio samples. A comparative evaluation of the Random Forest and CNN models is presented in Table III. The Random Forest classifier achieved an accuracy of 92.5%, with an AUC value of 0.94 and an F1-score of 0.91. Although this model provides relatively fast inference due to its lightweight structure, its performance is slightly lower than that of the deep learning model.

The CNN model achieved superior performance, reaching an accuracy of 96.8% and an AUC value of 0.98. The improved performance can be attributed to the CNN's ability to automatically learn discriminative spatial patterns within spectrogram representations of audio signals [6]. These patterns capture subtle artifacts introduced by synthetic speech generation methods that are difficult to detect using traditional handcrafted features.

TABLE III: PERFORMANCE COMPARISON

Model	Accuracy	AUC	F1-Score
Random Forest	92.5%	0.94	0.91
CNN	96.8%	0.98	0.96

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15377

TABLE IV: CONFUSION MATRIX

	Pred Real	**Pred Fake**
Actual Real	980	20
Actual Fake	45	955

The confusion matrix presented in Table IV further illustrates the classification performance of the CNN model. The model correctly classified the majority of both real and fake audio samples, with only a small number of misclassifications. Additionally, the receiver operating characteristic curve demonstrates a strong separation between the two classes, indicating a high discriminative capability of the proposed approach.

.

VIII. SYSTEM INTERFACE

The system includes real-time visualization support as shown in Fig. 4 and Fig. 5 for enhanced interpretability.

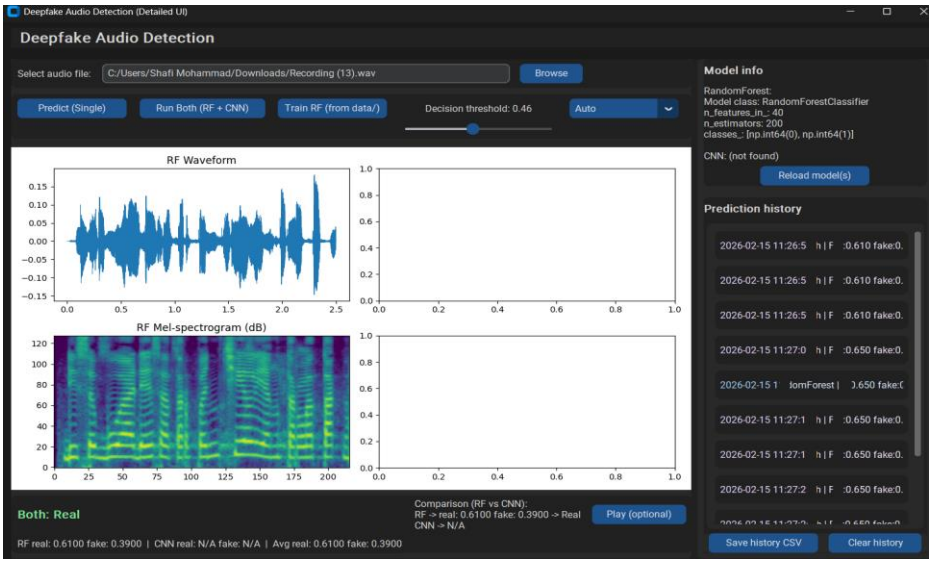

Fig. 4. Real Audio Prediction Output

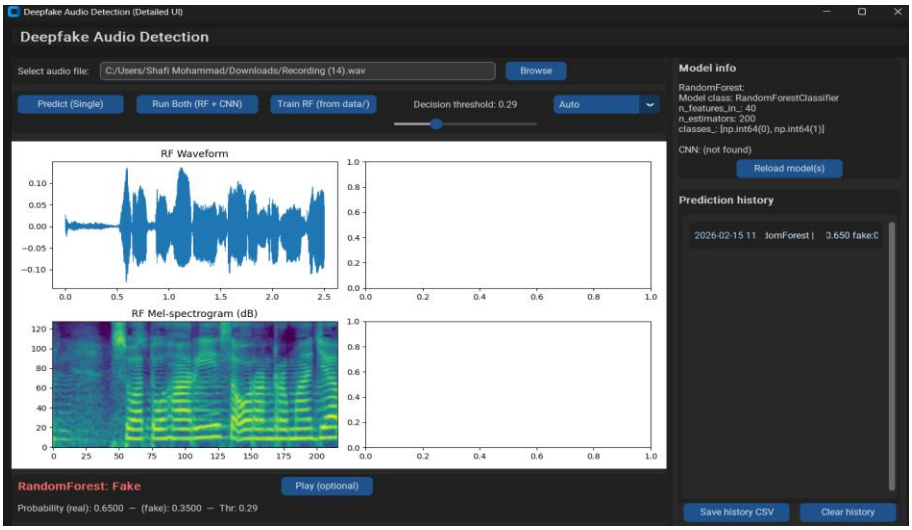

Fig. 5. Fake Audio Prediction Output

The "Run Both" mode enables comparative validation between RF and CNN outputs, helping users inspect consistency across the two classification approaches.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026
DOI: 10.17148/IJARCCE.2026.15377

IX. DISCUSSION

The hybrid system demonstrates that:

1) Classical models remain effective for lightweight inference.
2) CNN models provide superior accuracy.
3) Visual inspection aids interpretability.
4) Offline deployment enhances privacy.

Compared with deep end-to-end models such as RawNet [5], the proposed system requires fewer computational resources while maintaining strong performance. This makes the framework suitable for consumer-grade hardware and practical desktop deployment. The integration of machine learning and deep learning modules also improves flexibility, allowing users to choose between faster inference and higher predictive performance depending on the application scenario.

X. CONCLUSION

This paper presented a hybrid deepfake audio detection system combining Random Forest and Convolutional Neural Network models. The approach effectively balances computational efficiency and detection accuracy. Experimental results demonstrate high classification performance, with CNN achieving an accuracy of 96.8%. The developed desktop application enhances accessibility by integrating visualization and probability-based prediction. The system provides a practical solution for detecting synthetic speech and mitigating emerging audio-based threats. Future work will focus on improving model robustness against emerging neural speech synthesis techniques and evaluating the system on larger multilingual datasets.

REFERENCES

[1]. M. Todisco et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," Proc. Interspeech, 2019.
[2]. T. Kinnunen et al., "The ASVspoof 2017 Challenge," Proc. Interspeech, 2017.
[3]. J. Yamagishi et al., "ASVspoof 2021 Challenge," IEEE SLT, 2021.
[4]. M. Sahidullah et al., "A Comparison of Features for Synthetic Speech Detection," Proc. Interspeech, 2015.
[5]. H. Jung et al., "RawNet: End-to-End Neural Network Using Raw Waveforms," Proc. Interspeech, 2019.
[6]. G. Lavrentyeva et al., "Audio Spoofing Detection Using LCNN," Proc. Interspeech, 2019.
[7]. K. He et al., "Deep Residual Learning for Image Recognition," IEEE CVPR, 2016.
[8]. A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," arXiv:1609.03499, 2016.
[9]. Y. Choi et al., "StarGAN-VC: Many-to-Many Voice Conversion," IEEE SLT, 2018.
[10]. S. Davis and P. Mermelstein, "Comparison of Parametric Representations," IEEE TASSP, 1980.
[11]. I. Goodfellow et al., "Generative Adversarial Nets," NIPS, 2014.
[12]. D. Snyder et al., "X-vectors: Robust DNN Embeddings," ICASSP, 2018.
[13]. Y. Zhang et al., "Adversarial Learning for Speech Spoofing Detection," IEEE Access, 2020.
[14]. J. Chorowski et al., "Attention-Based Models for Speech," NIPS, 2015.
[15]. A. Gulati et al., "Conformer: Convolution-augmented Transformer," Proc. Interspeech, 2020.

