International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Deepfake Audio Detection Using Machine Learning

K. Sandeep
Department of Information Technology Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India

P. K. Abhilash
Department of Information Technology Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India

Raashi Sahu
Department of Information Technology Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India

Yarra Poojitha
Department of Information Technology Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India

Gundeti Bhargavi Reddy
Department of Information Technology Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India

***Abstract*—The fast growth of generative technologies has made synthetic, or deepfake, audio a significant threat to cybersecurity, digital fraud, and voice authentication. The proposed project is a machine learning method that separates real from artificially generated audio clips. Key audio features are generated through the extraction of acoustic measures using Librosa, as well as well-established measures such as Mel Frequency Cepstral Coefficients (MFCCs), chroma, and spectral contrast, that may allow for subtle changes in the quality of speech to be detected. A Random Forest classifier is then trained to perform binary classification of curated datasets of real and fake audio clips. To increase generalizability and robustness, noise-based augmentation is used during preprocessing. The proposed system achieves an accuracy of 97.8% and an F1-score of 98.7 on the SceneFake dataset. It is deployed using a Streamlit-based interface that allows users to upload their own audio files, view the waveform and spectrogram of the audio, and receive classifications with confidence scores in real-time. This proposed solution illustrates a lightweight, efficient, and practical overall framework for detecting deepfake audio without the complexity and computational overhead associated with deep learning models and processes, and becomes a feasible option for both real-time deployment and in low-resource limitations.**

***Keywords*—Deepfake Audio, Machine Learning, Audio Classification, MFCC Features, Librosa, Random Forest Classifier, Streamlit UI, Cybersecurity**

I. Introduction

In recent months, improvements in artificial intelligence and machine learning have vastly improved the field of speech synthesis and voice cloning. Deep learning frameworks can create very realistic sounding voices that are often nearly indistinguishable from real recordings. Such advancements are a significant step forward in a variety of domains, including entertainment, assistive technologies, and communications, while also presenting very real risk because deepfake audio can now be synthesized that closely imitate a real person's voice either playfully, or in a disingenuous manner, with the intention to impersonate or mislead. As such risks increase, there is a growing need for automated systems to quickly and accurately identify and distinguish manipulated voices from authentic recordings. Manual inspection methods are inadequate as the differences between real and synthetic voice recordings are often too subtle for the human ear to distinguish.

The Deepfake Audio Detection System tackles this challenge by leveraging audio signal processing and machine learning to accurately identify synthetic voices. First, the proposed system pulls out informative audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral contrast to characterize the unique spectral-temporal patterns of human speech. A Random Forest classifier is then trained with these features to predict whether audio is deepfake or real. To enhance accessibility and usability, a Streamlit-based web app is also built. This app allows users to upload an audio file, process it quickly, and outputs the results with clear visual indicators.

The Deepfake Audio Detection System aims to advance digital forensics and AI safety by combating the misuse of synthetic media, while demonstrating the role of machine learning in enhancing security and societal awareness. The key contributions of this work are: (1) a lightweight, feature based deepfake audio detection pipeline optimized for low resource environments; (2) an empirical evaluation showing competitive performance against more complex deep learning approaches; and (3) a real-time deployable Streamlit-based interface enabling practical forensic usage.

II. Literature Review

Recent studies on synthetic and deepfake speech detection have primarily focused on deep learning–based approaches. Todisco et al. [1] demonstrated that convolutional neural networks (CNNs) are effective in distinguishing real and vocoder-generated speech; however, their evaluation was limited to English datasets, restricting cross-lingual generalization. Feature-based methods have also been explored, such as the WaveFake framework proposed by Frank et al. [2], which utilizes spectrogram representations and support vector machines. While effective for earlier text-to-speech systems, the approach exhibits reduced performance when applied to newer synthesis techniques. Like modern, high-fidelity neural synthesis. Similarly, Reimao and Tzerpos [3] combined Mel-Frequency Cepstral Coefficients (MFCCs) with ensemble classifiers, achieving promising results under controlled conditions but showing performance degradation in noisy environments.

To address these robustness issues, Patel and Shah [4] explored data augmentation, finding that while it improves resilience to manipulation, excessive noise injection can degrade classification precision. More complex solutions have emerged, such as the hybrid CNN-RNN model proposed by Singh et al. [5], which captures both spectral and temporal dependencies at the cost of high computational requirements. Furthermore, Xiong et al. [6] utilized Transformer-based

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

contextual embeddings for detection, though these models often suffer from evaluation bias. Despite the high accuracy of deep learning approaches, their significant computational overhead limits their utility in real-time or resource-constrained applications. This creates a critical need for the lightweight, feature-based framework investigated in this study.

III. Methodologies

A. System Overview and Architecture

The detection system that is proposed has a modular pipeline that has been optimized to execute in a low latency. The pipeline has four distinct phases in order of execution: Signal Pre-processing; Multi-feature Extraction; Random Forest Classification; Deployment using Streamlit. In addition to the aforementioned phases of the pipeline, the feature driven approach allows the system to achieve a significant reduction in computational complexity compared to a purely end-to-end (E2E) deep learning approach which makes it the ideal candidate for the many organisations where CPU processing is the limiting factor and/or where the speed of analysis is critical in forensic investigations.

B. Data Collection and Pre-processing

Using the SceneFake dataset (which contains both real and synthetic speech, subjected to a variety of different acoustic distortions), researchers conducted several experiments through the creation of multiple training, development, and evaluation data subsets. Each of the dataset's samples are standardized based on a 16 kHz sample rate, and in a 16-bit RIFF/WAVE format. The dataset also contains many levels of Signal-to-Noise Ratio (SNR) ranging from -5 dB to 20 dB in order to simulate real-world scenarios. Pre-processing involves:

- Normalization: Amplitude scaling so that they have the same amount of signal power.
- Silence Removal: Trimming non-speech segments to focus the model on vocal characteristics.
- Data Augmentation: Addition of white Gaussian noise (AWGN) to enhance the adaptability of the model to unseen recording hardware.

C. Feature Extraction

Audio features are extracted using the Librosa library, to create a high-dimensional feature set of the signal. The extracted features include:

- Mel-Frequency Cepstral Coefficients (MFCCs): 20 coefficients are computed to represent the short-term power spectrum, mimicking human sound perception.
- Chroma Features: They capture energy distribution across the twelve pitch classes, to detect the harmonic inconsistencies that are commonly found in deepfake audio.
- Spectral Contrast: To highlight differences between spectral peaks and valleys to identify artifacts introduced by neural vocoders.

These feature are statistically summarized and concatenated to form fixed-length feature vectors for each audio signal.

D. Classification Model

A Random Forest classifier is used for binary classification of real versus synthetic audio. This model is selected for its ensemble learning capabilities and resistance to overfitting in high-dimensional spaces. The model is trained using labelled feature vectors and evaluated using a held-out test set. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess detection effectiveness. Unlike “black-box” deep learning models, the Random Forest architecture provides feature importance rankings which help in better interpretability of which acoustic cues are most indicative of deepfakes.

E. System Deployment and User Interface

The system is deployed via a Streamlit web application to bridge the gap between forensic research and practical utility. The interface provides:

- Visual Analytics: Real-time generation of waveforms and Mel-spectrograms for manual inspection.
- Inference Engine: A backend that processes uploaded audio files through the trained pipeline to return a binary label (Real/Fake).
- Probability Score: A confidence metric to indicate model’s certainty which is critical for decision-making in cybersecurity contexts.

IV. Results

A. Confusion Matrix Analysis

The performance of the Random Forest classifier was first evaluated using a confusion matrix to see how accurately the model distinguished between real and fake audio samples. The test dataset contained 981 real audio samples and 4,225 fake audio samples.

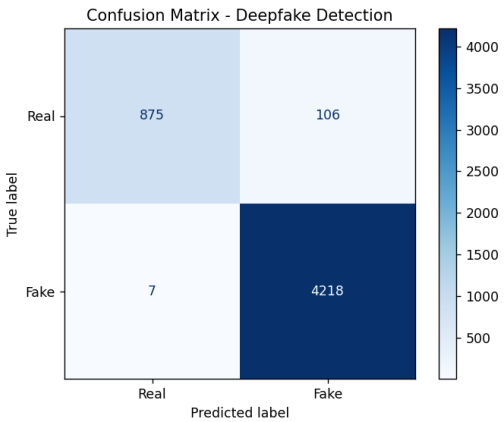

Fig. 1. Confusion Matrix

As illustrated in Fig. 1, the model correctly identified 875 real samples and 4218 fake samples, only 7 fake audio samples were misclassified as real.

In context of cybersecurity, a low False Negative rate is excellent as the failure to detect a deepfake has higher risk that a false alarm on an authentic recording.

B. Model Evaluation Metrics

To get a detailed quantitative assessment, the model was evaluated using standard classification metrics. Despite the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

class imbalance in the SceneFake dataset, the Random Forest Classifier had fair robustness:

- Accuracy: 97.8%
- Precision: 97.5%
- Recall: 99.8%
- F1-Score: 98.7%

The evaluation metrics confirm that the proposed Random Forest model is both accurate and reliable in distinguishing real from synthetic audio. The high recall value shows that nearly all deepfake samples are detected. Although the dataset is slightly imbalanced, the high recall and precision indicate that the model maintains reliable performance across classes. These results further validate that a lightweight, feature-based approach can achieve performance similar to computationally expensive deep learning models while remaining feasible for low-resource deployment.

C. System Demonstration and UI Integration

The proposed framework was demonstrated using a Streamlit-based user interface developed to evaluate its practical usability. The application supports real-time inference and provides multiple visual representations of input audio signal to assist in analysis.

User Interface (Fig. 2): The interface consists of a simple and user-friendly dashboard that allows users to upload audio files quickly.

Input and Signal Visualization (Fig. 3): After user uploads the audio file, the system generates an audio waveform so user can observe the temporal structure of the speech signal and to identify any irregular patterns.

Spectogram and Prediction (Fig. 4): In addition to waveform, the spectrogram is displayed along with the final classification output. The spectrogram enables forensic analysts to observe spectral patterns like checkerboard artifacts or high-frequency noise which are common in synthetic audio. The system also gives confidence score along with the predicted label to show reliability of the predicted result.

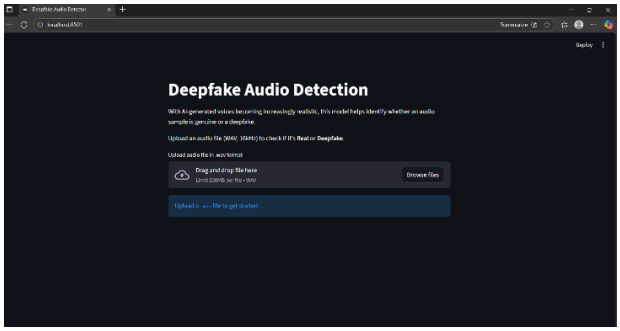

Fig. 2. User Interface

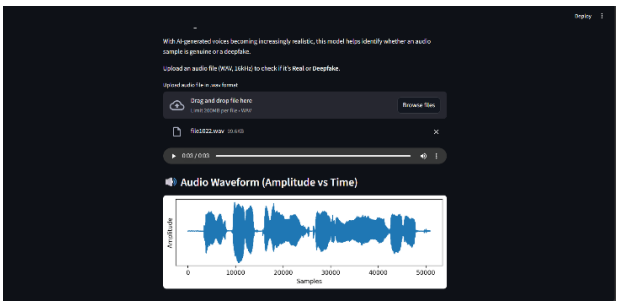

Fig. 3. Input

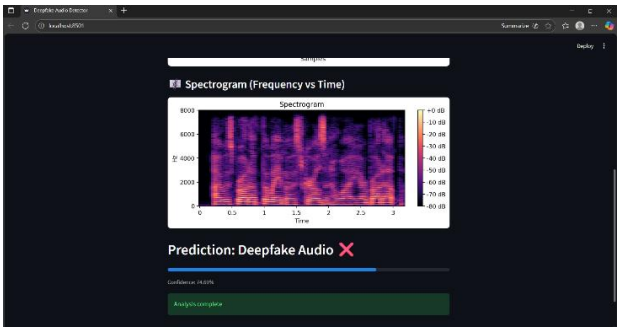

Fig. 4. Output

V. Conclusion

The rapid evolution of voice cloning technologies has created a critical need for reliable detection tools that can protect the integrity of digital communications. This paper presented a practical approach to this problem by prioritizing a lightweight, feature-driven machine learning pipeline over computationally heavy deep learning models. By targeting specific audio features like MFCCs, Chroma, and Spectral Contrast, we developed a system that effectively captures the subtle "digital fingerprints" left by synthetic voice generators.

Our experimental results on the SceneFake dataset are highly encouraging, with a 97.8% accuracy rate and a 99.8% recall. The high recall is particularly vital in a security context, as it ensures that virtually no synthetic attempts bypass the system. Furthermore, the development of the Streamlit interface moves this research from a theoretical exercise into a usable tool, providing real-time visual and probabilistic feedback. Ultimately, this work demonstrates that by focusing on high-quality feature engineering and ensemble learning, we can create forensic tools that are not only accurate but also efficient enough for immediate, real-world deployment.

VI. Future Prospects

A. Hybrid Modelling

While our current Random Forest model is highly efficient, we plan to experiment with hybrid architectures. Integrating lightweight CNN layers could help the system learn more complex spectral patterns, further improving performance against high-fidelity neural vocoders.

B. Broadening Dataset Diversity

Synthetic voices behave differently across different languages and accents. We intend to expand our training sets to include a more global range of dialects and diverse

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

background noise conditions (such as street noise or office environments) to ensure the model generalizes well across all user demographics.

C. *Live Stream Processing*

A major goal is to move beyond file-based uploads and implement a "streaming" analysis mode. This would allow the system to monitor live audio feeds or phone calls, providing an active layer of defense against real-time impersonation attacks.

D. *Scalability via Cloud and Mobile*

To reach a wider audience, we are looking into converting the Streamlit prototype into a full-scale web API and a mobile application. Hosting the model on cloud platforms would allow developers to integrate our "Deepfake-Detection-as-a Service" into existing communication apps and authentication workflows.

ACKNOWLEDGMENT

The authors would like to thank Mr. K. Sandeep and Mr. P. K. Abhilash, Associate Professors at Gokaraju Rangaraju Institute of Engineering and Technology, for their valuable guidance and insightful suggestions throughout the development of this work.

REFERENCES

[1] A. Todisco, H. Delgado, and N. Evans , “Detection of synthetic speech using deep learning techniques,” in Proc. Interspeech, 2019, pp. 1008 1012.

[2] J. Frank, A. Schönherr, L. Specht, and T. Holz, “WaveFake: A data driven explainable approach to detecting speech deepfakes,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2020, pp. 6573-6577.

[3] A. Reimao and T. Tzerpos, “Combining MFCCs and ensemble classifiers for spoofed speech detection,” IEEE Access, vol. 9, pp. 165604-165615, 2021.

[4] K. Patel and P. Shah, “Improving deepfake audio detection using data augmentation techniques,” in Proc. IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR), 2023, pp. 45-50.

[5] R. Singh, A. Kaur, and S. Bansal, “Hybrid CNN–RNN framework for deepfake speech pattern recognition,” in Proc. Int. Conf. Artif. Intell. Signal Process. (AISP), 2024, pp. 212-217.

[6] W. Xiong, Y. Lin, and M. Chen, “Transformer-based architectures for synthetic voice detection,” in Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), 2022, pp. 1482-1487.