Deepfake Audio Detection with Neural Networks Using AI and Audio Features

Aiswariya A[1], Abisha J[2], Adithya S Kumar[3], Dhivya V M[4]
[1,2,3]*UG Student, Department of Computer Science & Engineering, Sivaji College of Engineering and Technology*
[4]*Assistant Professor, Department of Computer Science and Engineering, Sivaji College of Engineering and Technology*
doi.org/10.64643/IJIRTV12I11-203610-459

***Abstract*—AI-generated deepfake audio poses serious risks to security, misinformation, and identity fraud, creating a need for reliable detection systems. This project develops a deepfake audio detection framework using spectral and temporal voice features, voiceprint analysis, and AI-based classification models. The system combines Convolutional Neural Networks (CNN) for feature extraction and Bi-LSTM networks for temporal pattern analysis to accurately distinguish synthetic speech from real human voices.**

A real-time filtering and probabilistic scoring mechanism further improves verification and detection reliability. Trained on diverse datasets with multiple speech synthesis techniques, the model analyzes pitch, tone, and spectral inconsistencies to provide a scalable and effective solution against AI-driven audio manipulation.

Existing deepfake detection techniques often struggle to adapt to evolving speech synthesis methods and may fail to identify subtle inconsistencies in audio signals. Therefore, this project focuses on developing a robust and adaptive system using Convolutional Neural Networks (CNN) and Bi-LSTM networks to analyze both spectral and temporal characteristics of speech. In addition, the integration of real-time filtering and probabilistic scoring mechanisms aims to provide accurate, fast, and reliable verification suitable for real-world applications such as media authentication, voice-based security systems, and online communication platforms.

Attackers can create convincing fake audio recordings to impersonate public figures, company executives, or ordinary individuals for malicious purposes. Traditional authentication and security systems are not designed to detect such advanced manipulations, highlighting the need for intelligent and automated detection frameworks.

I. INTRODUCTION

The rapid growth of artificial intelligence has enabled the creation of highly realistic synthetic speech, raising concerns about security, trust, and authenticity in digital communication. Deepfake audio can be misused for identity theft, fraud, and misinformation, making reliable detection systems essential. Existing methods often fail to keep pace with advanced audio manipulation techniques, creating the need for more adaptive solutions.

This project aims to develop an intelligent deepfake audio detection framework using CNN and Bi-LSTM models to analyze spectral and temporal speech patterns. The system also incorporates real-time verification and probabilistic scoring to provide accurate, fast, and practical detection of manipulated audio in real-world applications

II. PROBLEM STATEMENT

The rapid development of artificial intelligence and deep learning technologies has led to the creation of highly realistic deepfake audio that can closely imitate human speech. While synthetic voice generation has useful applications in entertainment, accessibility, and virtual assistance, it also introduces serious risks related to misinformation, identity theft, fraud, and cybercrime. Deepfake audio can be used to manipulate public opinion, impersonate individuals, and bypass voice-based authentication systems, creating major concerns regarding trust, privacy, and digital security. Existing audio authentication and detection methods are often inadequate in identifying advanced synthetic speech generated by modern AI models. Many

traditional approaches fail to capture subtle spectral and temporal inconsistencies present in manipulated audio, especially as deepfake generation techniques continue to evolve. In addition, several existing systems lack adaptability, robustness, and real-time processing capabilities, making them unsuitable for practical deployment in security-sensitive environments such as banking, media verification, and online communication platforms.

To address these challenges, there is a need for an intelligent and efficient deepfake audio detection framework capable of accurately distinguishing between genuine and AI-generated speech. This project proposes the use of Convolutional Neural Networks (CNN) for feature extraction and Bi-LSTM networks for temporal sequence analysis to improve detection accuracy.

III. OBJECTIVES

The objective of this project is to develop an intelligent deepfake audio detection system capable of accurately distinguishing between real and synthetic speech using CNNs for feature extraction and Bi-LSTM networks for temporal analysis. The system aims to assign authenticity scores, automate verification, and reduce false positives through real-time detection mechanisms. Additionally, the model is trained on diverse datasets and evaluated using performance metrics such as precision, recall, and F1-score to ensure reliability, accuracy, and robustness in real-world applications. The proposed framework also focuses on improving cybersecurity and preventing misuse of AI-generated voice cloning technologies.

IV. LITERATURE REVIEW

The literature survey focuses on studying existing research and techniques used for deepfake audio detection to understand their strengths, limitations, and effectiveness. It helps identify research gaps and supports the selection of suitable methods for developing an improved and accurate deepfake detection framework. Various studies have explored the use of deep learning, audio feature extraction, and neural network models to distinguish between genuine and manipulated speech.

A study by IEEE titled “Deepfake Audio Detection with Neural Networks Using Audio Features” proposed a CNN-based approach that converts audio signals into image-like representations such as Spectrograms, MFCC, FFT, and STFT for efficient feature extraction. This method reduces computational complexity and improves classification between human and synthetic speech. However, the system may still face challenges in handling advanced spoofing attacks and requires large datasets for improved generalization.

Another important contribution is the review paper “Speech Recognition Using Deep Neural Networks: A Systematic Review”, which provides a detailed analysis of deep learning applications in speech processing. The study highlights the effectiveness of CNNs, LSTMs, and other neural network models in speech recognition and audio analysis. Similarly, surveys on deepfake detection techniques using deep learning discuss the rapid growth of AI-generated media and summarize state-of-the-art detection methods. These studies provide valuable insights into feature extraction, model architectures, and challenges in detecting increasingly sophisticated deepfakes.
Research on multimodal deepfake detection has further improved detection accuracy by combining both audio and visual information. The paper “Emotions Don’t Lie: An Audio-Visual Deepfake Detection Method using Affective Cues” introduced a deep learning approach that analyzes emotional consistency between audio and video signals to identify fake content. The proposed model achieved high accuracy on benchmark datasets such as DeepFake-TIMIT and DFDC

V. EXISTING SYSTEM

The existing deepfake audio detection systems mainly rely on traditional machine learning techniques such as Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) for classifying audio signals as real or fake. These systems use handcrafted audio features like Mel-Frequency Cepstral Coefficients (MFCC), pitch, spectral centroid, chroma features, and zero-crossing rate to analyze speech patterns. These methods provided reasonable performance in detecting basic audio manipulations and replay attacks.

The existing deepfake audio detection systems mainly rely on traditional machine learning techniques such as Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) for classifying audio signals as real or fake. These systems use handcrafted audio features like Mel-Frequency Cepstral Coefficients (MFCC), pitch, spectral centroid, chroma features, and zero-crossing rate to analyze speech patterns.

Feature extraction plays a major role in these approaches, where audio signals are converted into structured numerical representations before classification. These methods provided reasonable performance in detecting basic audio manipulations and replay attacks. Despite their advantages, the existing systems face several limitations when dealing with highly realistic deepfake audio generated using advanced AI models such as GANs, WaveNet, and transformer-based architectures. These systems often fail to detect advanced AI-generated audio accurately due to limited feature analysis.

VI. PROPOSED SYSTEM

The proposed system is designed to detect deepfake audio accurately by combining advanced deep learning techniques with intelligent audio analysis. The system first performs preprocessing operations such as noise removal, normalization, and signal enhancement to improve audio quality. Important speech features such as MFCC, pitch, tone, and spectral characteristics are then extracted from the input audio. These features help represent the unique properties of speech and provide meaningful input for the deep learning model.

The core of the proposed system is a hybrid deep learning architecture that combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks. CNN is used for extracting spatial and spectral features from spectrogram images, while Bi-LSTM captures temporal dependencies and sequential speech patterns. This hybrid approach enables the system to identify subtle inconsistencies and manipulations present in AI-generated audio. The model performs binary classification to determine whether the input speech is genuine or deepfake and generates confidence scores for reliable decision-making.

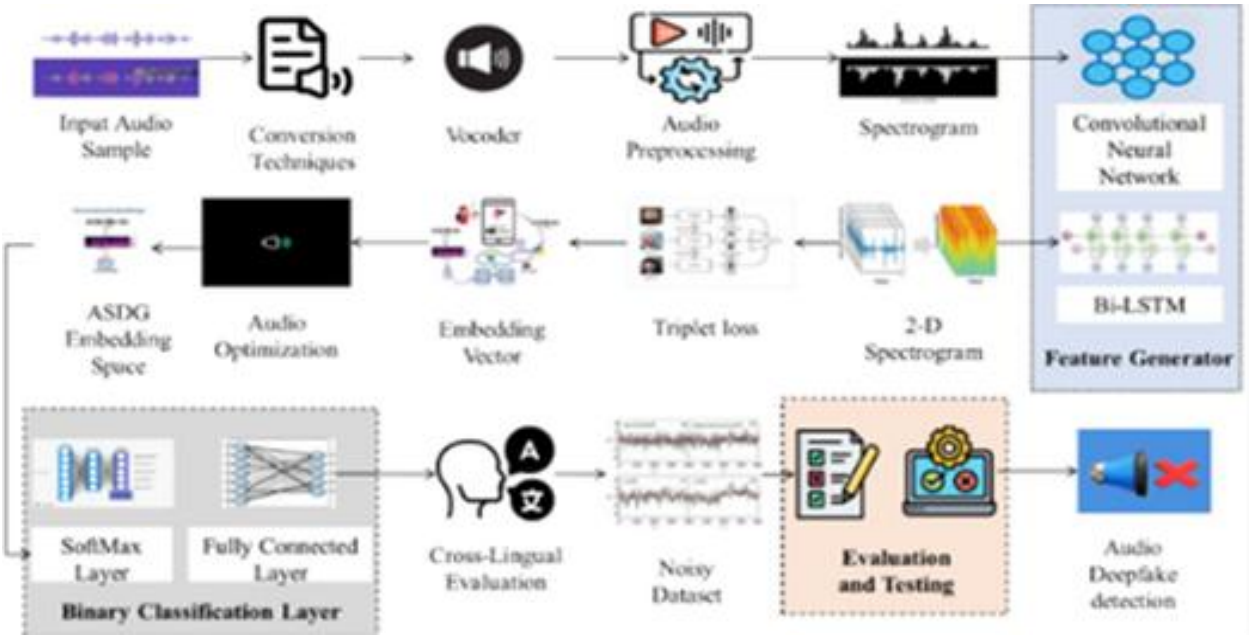
The proposed framework also includes real-time verification and repeat content detection mechanisms to improve system security and robustness. If suspicious or AI-generated speech is detected, the system automatically triggers alerts and prevents unauthorized access. The model is trained on diverse datasets to improve adaptability and generalization across different deepfake generation techniques. By integrating automated detection, probabilistic scoring, and real-time monitoring, the proposed system provides a scalable, efficient, and reliable solution for protecting voice-based applications from AI-driven audio manipulation.

VII. SYSTEM ARCHITECTURE

The system architecture of the proposed deepfake audio detection system consists of multiple stages designed to analyze and classify audio signals as real or AI-generated. The process begins with audio input collection, where voice data is obtained from recorded files or real-time sources. The input audio then passes through a preprocessing stage that performs noise removal, normalization, and signal enhancement to improve audio quality. After preprocessing, the audio is converted into suitable representations such as spectrograms and MFCC features, which help capture important frequency and time-based speech characteristics required for accurate analysis. The extracted features are then fed into a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks.

If suspicious or AI-generated speech is detected, the system automatically triggers alerts and prevents unauthorized access. The model is trained on diverse datasets to improve adaptability and generalization across different deepfake generation techniques. By integrating automated detection, probabilistic scoring, and real-time monitoring, the proposed system provides a scalable, efficient, and reliable solution.

ARCHITECTURE DIAGRAM



The fig. of system architecture for deepfake audio detection

VIII. MODULES DESCRIPTION

The Audio Input Module is responsible for collecting voice samples from recorded files or real-time sources. It accepts both genuine and suspicious audio signals for further analysis. This module ensures that the audio is properly captured and forwarded to the next processing stage without data loss.

The Preprocessing Module improves the quality of the input audio by performing noise removal, normalization, and signal enhancement. It removes unwanted disturbances and converts the audio into a clean and standardized format. This helps improve the accuracy and efficiency of the detection system.

The Feature Extraction Module extracts important speech features such as MFCC, pitch, tone, and spectral features from the audio signal. These features represent the unique characteristics of speech and help distinguish between real and fake audio. The extracted features are converted into numerical representations.

The Convolutional Neural Network (CNN) Module analyzes spectrogram images and extracts important spatial and spectral patterns from the audio. It identifies hidden inconsistencies, frequency variations, and manipulation artifacts present in synthetic speech. This module improves the system's ability to detect deepfake audio accurately.

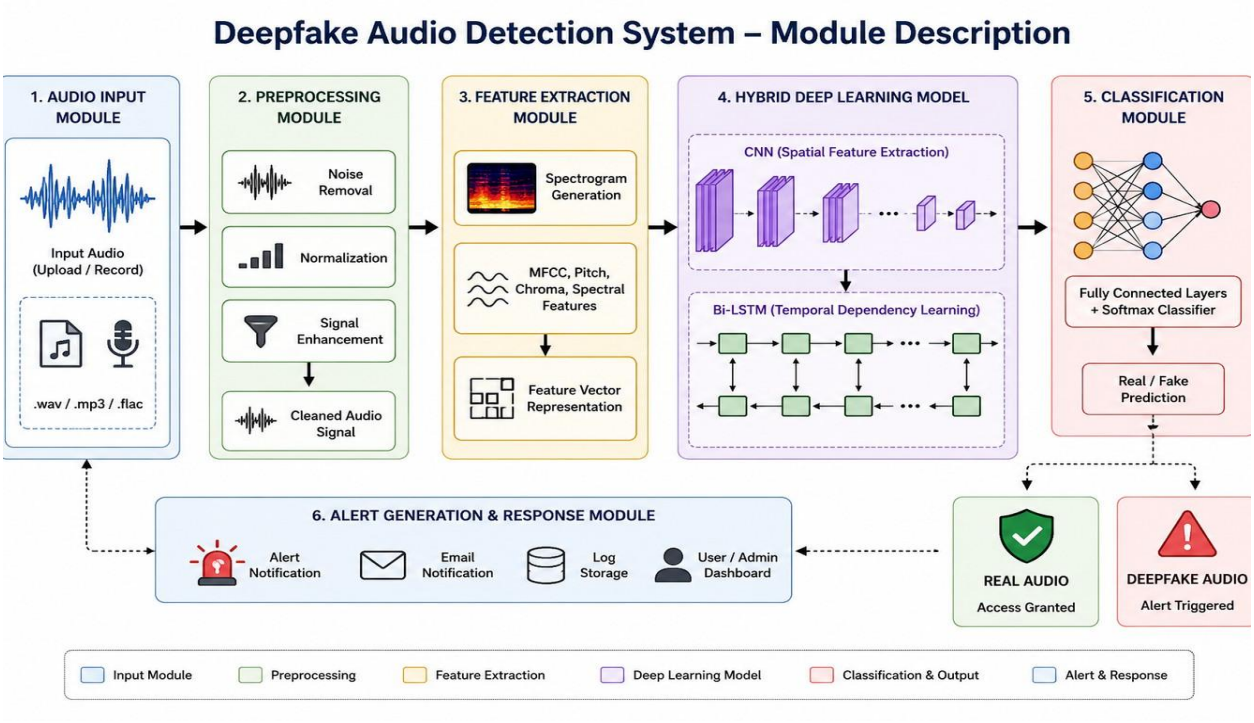
The Bidirectional Long Short-Term Memory (Bi-LSTM) Module captures temporal dependencies and sequential speech patterns from the audio signal. It processes speech data in both forward and backward directions to better understand variations in rhythm, pauses, and tone. This helps in identifying advanced AI-generated speech manipulations.

The Classification and Alert Module classifies the audio as Real or Deepfake using confidence scores generated by the model. If suspicious audio is detected, the system automatically triggers alerts or notifications for security purposes. This module enhances reliability, authentication, and real-time fraud prevention in voice-based applications.

The proposed deepfake audio detection system is designed to identify whether an audio sample is genuine or artificially generated using advanced deep learning techniques. The system begins by accepting audio input from uploaded recordings or live voice samples. The received audio is then processed using preprocessing techniques such as noise removal, signal enhancement, and normalization to improve audio quality and remove

unwanted disturbances. Clean and standardized audio signals help improve the reliability and performance of the detection framework.

After preprocessing, the system performs feature extraction to capture important speech characteristics from the audio signal. Features such as MFCC, pitch, tone, spectral features, and spectrogram representations are extracted and converted into numerical formats suitable for machine learning analysis. These extracted features contain valuable information related to speech frequency, intensity, and temporal variations, which are useful for distinguishing between natural and synthetic voices.



The fig. of Modules description

IX. RESULTS

The experimental results of the proposed Deepfake Audio Detection System demonstrate that the hybrid deep learning model effectively distinguishes between genuine and AI-generated speech. The system was trained and tested using diverse audio datasets containing both real and synthetic voice samples. During evaluation, the model successfully identified hidden spectral and temporal inconsistencies present in manipulated audio. The combination of Convolutional Neural Networks (CNN) and Bi-LSTM networks significantly improved feature learning and sequence analysis, resulting in better detection accuracy compared to traditional machine learning approaches.

The performance of the model was measured using important evaluation metrics such as accuracy, precision, recall, and F1-score. The results showed that the proposed framework achieved high classification performance while minimizing false positives and false negatives. CNN efficiently extracted spatial and spectral patterns from spectrograms, whereas Bi-LSTM captured temporal dependencies and speech flow variations. The hybrid architecture improved the system's ability to detect even high-quality synthetic speech generated using advanced AI techniques. The probabilistic scoring mechanism also enhanced the reliability of prediction results.

X. FUTURE ENHANCEMENT

In the future, the proposed deepfake audio detection system can be enhanced by integrating more advanced deep learning models such as Transformers, Attention Mechanisms, and Generative Adversarial Network (GAN) based detectors to improve detection accuracy

against highly sophisticated synthetic speech. The framework can also be trained on larger and multilingual datasets to improve adaptability across different languages, accents, and speaking styles. Incorporating advanced voice biometrics and behavioral analysis can further strengthen the system's capability to detect subtle manipulations in AI-generated audio.

Another important enhancement is the development of a fully automated real-time monitoring and cloud-based deployment system for large-scale applications. The framework can be integrated with voice authentication systems, social media platforms, and cybersecurity applications to provide continuous protection against deepfake misuse. Future improvements may also include mobile application support, live call monitoring, and explainable AI techniques to provide transparent and interpretable detection results. These enhancements will make the system more scalable, secure, and efficient for real-world deployment.

XI. CONCLUSION

The proposed Deepfake Audio Detection System provides an effective solution for identifying AI-generated and manipulated speech using advanced deep learning techniques. By combining Convolutional Neural Networks (CNN) and Bi-LSTM networks, the system successfully analyzes both spectral and temporal speech characteristics to distinguish between genuine and synthetic audio. The integration of preprocessing, feature extraction, and automated classification improves the overall accuracy and reliability of detection.

The system also supports real-time verification and alert generation, making it suitable for practical applications such as voice authentication, cybersecurity, and media verification. Experimental results demonstrate that the framework achieves high performance in detecting deepfake audio while reducing false predictions. Overall, the proposed model offers a scalable, secure, and efficient approach.

REFERENCES

[1] W. Al-Dulaimi, T. K. Moon, and J. H. Gunther, "Voice transformation using two-level dynamic warping and neural networks," *Signals*, vol. 2, no. 3, pp. 456–474, 2021, doi: 10.3390/signals2030026.

[2] M. Almars, "Deepfakes detection techniques using deep learning: A survey," *Journal of Computer and Communications*, vol. 9, no. 5, pp. 20–35, 2021, doi: 10.4236/jcc.2021.95002.

[3] T. Balamurali, K. E. Lin, S. Lui, J. M. Chen, and D. Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84229–84241, 2019, doi: 10.1109/ACCESS.2019.2924383.

[4] X. Chen, Y. Zhang, G. Zhu, and Z. Duan, "UR channel-robust synthetic speech detection system for ASVspoof 2021," *arXiv preprint arXiv:2107.12018*, 2021.

[5] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, "A study on data augmentation in voice anti-spoofing," *Speech Communication*, vol. 141, pp. 56–67, 2022, doi: 10.1016/j.specom.2022.04.002.

[6] W. Ge, J. Patino, M. Todisco, and N. Evans, "Raw differentiable architecture search for speech deepfake and spoofing detection," *arXiv preprint arXiv:2107.12212*, 2021.

[7] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 8481–8490, doi: 10.1109/CVPR42600.2020.00850.