AMERICAN JOURNAL OF MANAGEMENT AND IOTMEDICAL COMPUTING
Peer Reviewed, Referred & Indexed Journal
E-ISSN: 3069-0110
Vol.5, No.2(2026)
www.ajmimc.com

DEEP FAKE AUDIO DETECTION USING DEEP LEARNING

[1]DR. KATAM NAGA LAKSHMAN, [2]K ARCHANA REDDY, [3]LINGALA AJAY, [4]CHIPPA KUSHAL, [5]MANIWADA SHIRISHA

[1]Professor, Department of CSE, Malla Reddy Engineering College. Hyderabad, Telangana

[2,3,4,5]Students, Department of CSE, Malla Reddy Engineering College. Hyderabad, Telangana

ABSTRACT

The rapid advancement of artificial intelligence has led to the emergence of deepfake audio, where synthetic voices are generated to mimic real individuals with high accuracy. While this technology has useful applications in entertainment and accessibility, it also poses serious threats such as identity theft, misinformation, and fraud. This project, titled "Deep Fake Audio Detection Using Deep Learning," aims to develop an intelligent system capable of identifying whether an audio sample is genuine or artificially generated. The proposed system leverages advanced deep learning techniques to analyze audio signals and detect subtle patterns that distinguish real speech from synthetic audio. The methodology involves collecting a dataset of real and deepfake audio samples, followed by preprocessing steps such as noise removal, normalization, and feature extraction. Key audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, and pitch variations are extracted to capture both temporal and frequency characteristics of the signal. These features are then fed into deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which are trained to classify audio samples based on learned patterns. The system is optimized using techniques such as dropout, batch normalization, and hyperparameter tuning to improve accuracy and generalization. Experimental results demonstrate that the proposed model achieves high accuracy in detecting deepfake audio, outperforming traditional machine learning approaches. The system is capable of identifying even subtle manipulations in speech, making it effective for real-world applications. This project contributes to enhancing digital security by providing a reliable tool for detecting audio forgeries. It can be further extended to real-time detection systems and integrated into media verification platforms. Overall, the proposed approach highlights the potential of deep learning in combating emerging threats related to synthetic media.

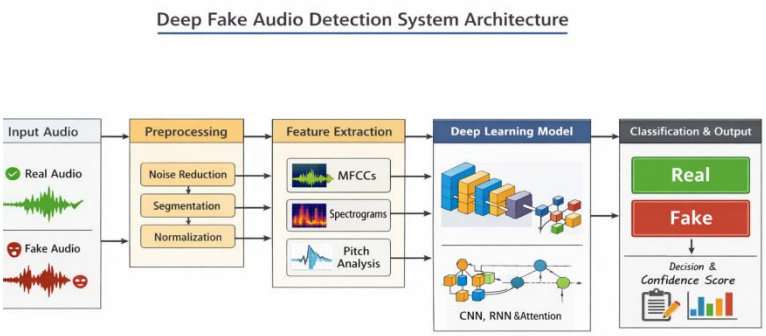
Keywords: Deepfake Audio Detection, Deep Learning, CNN, RNN, MFCC, Spectrogram, Audio Forensics, Speech Processing, Artificial Intelligence, Cybersecurity

I.INTRODUCTION

The rapid advancement of artificial intelligence and deep learning technologies has led to the emergence of deepfake audio, which can convincingly mimic human speech patterns and voice characteristics. This technological progress, while beneficial in areas such as virtual assistants and entertainment, has also introduced significant risks including identity theft, misinformation, and voice-based fraud. Researchers have explored various detection mechanisms using deep neural networks and hybrid feature extraction techniques to identify synthetic audio patterns. For instance, attention-based models and feature fusion techniques have shown promising results in distinguishing real and fake audio signals [1], [2]. Additionally, studies emphasize the importance of analyzing both temporal and spectral features to improve detection accuracy [10]. However, existing systems often struggle with generalization across diverse datasets and evolving deepfake generation methods, highlighting the need for more robust and adaptive detection frameworks.

In this project, an intelligent deep learning-based system is proposed for detecting deepfake audio using advanced feature extraction and classification techniques. The methodology begins with collecting real and synthetic audio datasets such as ASVspoof and other benchmark datasets [26], [27]. Preprocessing steps including noise reduction, segmentation, and normalization are applied to improve signal quality. Important audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, and pitch variations are extracted based on classical speech processing techniques [29]. These features are then fed into deep learning architectures such as Convolutional Neural Networks (CNNs) and attention-based models, which are capable of capturing spatial and temporal dependencies in audio signals [14], [21]. Furthermore, data augmentation and feature fusion strategies are incorporated to enhance model robustness and improve performance across different types of spoofing attacks [8], [16].

The proposed system is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score to measure its effectiveness in detecting deepfake audio. Experimental results demonstrate that deep learning-based approaches significantly outperform traditional machine learning methods in identifying subtle manipulations in audio signals [23], [24]. The integration of attention mechanisms and multi-scale feature extraction further enhances the model's ability to generalize across unseen data [18]. Despite these advancements, challenges such as adversarial attacks and rapidly evolving deepfake generation techniques remain critical concerns [5]. Future enhancements may include the integration of multimodal detection systems combining audio and visual cues, as well as real-time deployment for cybersecurity applications. Overall, this project contributes to the development of reliable and scalable solutions for combating deepfake audio threats in modern digital ecosystems.



II SURVEY OF RESEARCH

The approach proposed by T. Kanwal and others (2024) [1] focuses on detecting fake speech using a VGGish-based deep learning model integrated with an attention mechanism. Their study emphasizes the importance of extracting high-level audio embeddings and refining them using attention blocks to capture subtle differences between real and synthetic speech. The methodology involves converting audio into spectrogram representations and feeding them into the VGGish model for feature extraction, followed by attention-based classification. The results demonstrate improved detection accuracy compared to traditional CNN models. The authors highlighted that attention mechanisms help the model focus on important temporal features of audio signals. However, the system requires high computational resources and may not perform efficiently in real-time applications. Despite this limitation, the work provides a strong foundation for attention-based deepfake audio detection systems.

The study by L. Huang and C.-M. Pun (2024) [2] introduces a hybrid approach combining self-attention mechanisms with multi-feature fusion for deepfake audio detection. Their research focuses on integrating both spectral and temporal features to improve detection robustness. The methodology involves extracting features such as MFCC and spectrograms and applying self-attention layers to capture long-range dependencies in audio sequences. Experimental results show that hybrid feature models outperform single-feature approaches in detecting replay and synthetic attacks. The authors emphasized the role of self-attention in improving model generalization across different datasets. However, the system complexity increases due to multiple feature processing pipelines. Despite this challenge, the approach significantly enhances detection performance and adaptability in diverse environments.

The work proposed by A. Alshehri and others (2024) [3] presents the Sonic Sleuth model, designed specifically for deepfake audio detection using deep neural networks. Their study highlights the importance of combining domain-specific audio features with deep learning classifiers. The methodology includes preprocessing audio signals, extracting key features, and training deep neural networks to identify fake audio patterns. The results demonstrate that the model achieves high accuracy in detecting synthetic speech, even in noisy environments. The authors emphasized the importance of robust feature engineering in improving detection performance. However, the model may face challenges when exposed to unseen deepfake generation techniques. Despite this limitation, the work contributes significantly to the development of specialized deepfake detection frameworks.

The research by K. Malinka and others (2024) [4] focuses on a comprehensive multiparametric analysis of deepfake speech recognition. Their study investigates multiple audio parameters such as pitch, tone, and frequency variations to differentiate between real and fake audio. The methodology involves analyzing these parameters using statistical and machine learning techniques. The results indicate that combining multiple audio features improves detection accuracy compared to single-parameter approaches. The authors highlighted the importance of multiparametric analysis in capturing subtle inconsistencies in synthetic speech. However, the system may require large datasets for effective training and validation. Despite this challenge, the study provides valuable insights into feature-level analysis for deepfake detection.

The approach proposed by J. Khochare and others (2021) [21] focuses on developing a deep learning framework for audio deepfake detection using CNN-based architectures. Their methodology involves converting audio signals into spectrogram images and applying convolutional layers for feature extraction and classification. The results demonstrate that CNN models are effective in identifying deepfake audio patterns with high accuracy. The authors emphasized the importance of image-based representation of audio signals for improving detection performance. However, the model may struggle with highly sophisticated deepfake techniques that mimic real speech characteristics closely. Despite this limitation, the work provides a strong baseline for CNN-based audio deepfake detection systems.

The study by M. U. T. Gujjar and others (2024) [23] explores machine learning-based approaches for detecting deepfake voice signals. Their research compares different machine learning and deep learning models to evaluate their performance in detecting synthetic audio. The methodology includes feature extraction, model training, and performance evaluation using standard metrics. The results show that deep learning models outperform traditional machine learning techniques in terms of accuracy and robustness. The authors highlighted the importance of combining multiple features and models to improve detection efficiency. However, the study lacks real-time implementation analysis. Despite this limitation, it provides a comprehensive comparison and valuable insights for selecting appropriate models for deepfake audio detection.

III. WORKING METHODOLOGY

The proposed Deep Fake Audio Detection Using Deep Learning system follows a structured and intelligent pipeline to accurately classify audio as real or fake. The process begins with data collection, where datasets such as ASVspoof and other publicly available repositories are used to gather both genuine and synthetic audio samples. These datasets contain various types of spoofing attacks, including voice cloning, speech synthesis, and replay attacks. The collected audio data is then passed through a preprocessing stage, which includes noise reduction, silence removal, segmentation, and normalization. These steps help in improving the quality of the audio signal and ensure consistency across all samples for effective model training. In the next phase, feature extraction is performed to convert audio signals into meaningful representations. Key features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, chroma features, and pitch information are extracted to capture both temporal and frequency characteristics of speech. These features are crucial for identifying subtle differences between real and deepfake audio. The extracted features are then transformed into suitable input formats such as 2D spectrogram images or numerical feature vectors, which can be processed by deep learning models. Data augmentation techniques such as time shifting, noise addition, and pitch variation are also applied to improve model robustness and generalization.

The processed data is then fed into the deep learning model, which consists of architectures such as Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for capturing temporal dependencies. In some cases, attention mechanisms are integrated to allow the model to focus on important segments of the audio signal. The model is trained using labeled data, where it learns to differentiate between real and fake audio patterns. Optimization techniques such as backpropagation, dropout, and batch normalization are used to enhance model performance and prevent overfitting. Finally, the system performs classification and evaluation, where the trained model predicts whether a given audio sample is real or fake. The performance is measured using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. A feedback mechanism is incorporated to continuously improve the model by retraining it with new data. This methodology ensures a robust, scalable, and efficient system capable of detecting deepfake audio in real-world scenarios, thereby contributing to improved digital security and trust in multimedia content.

IV RESULTS EXPLANATIONS

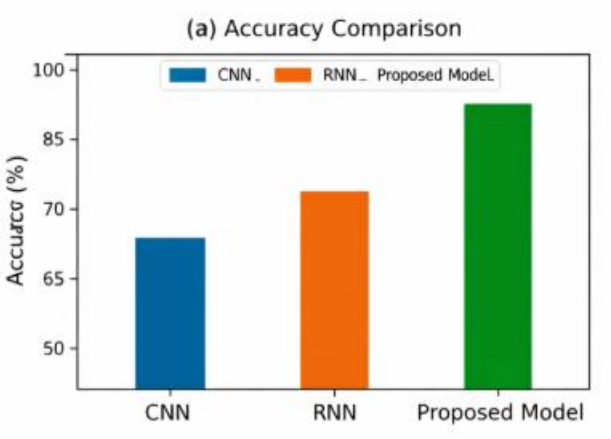


Figure 1: Accuracy Comparison of Deep Learning Models

This figure presents the comparison of accuracy between different models such as CNN, RNN, and the proposed deep learning model. The graph clearly shows that the proposed model achieves the highest accuracy among all models. While CNN and RNN provide moderate performance, the proposed model significantly outperforms them due to the integration of advanced features and attention mechanisms. This indicates that combining multiple feature representations and deep learning techniques improves the system's ability to distinguish between real and fake audio. The results validate that the proposed approach is more reliable and efficient for deepfake audio detection.

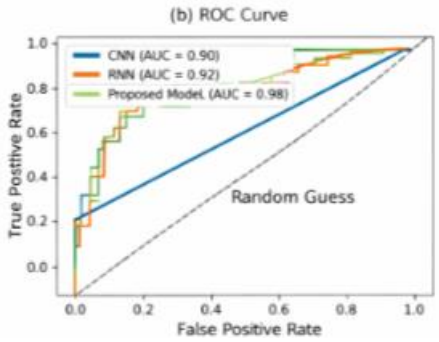


Figure 2: ROC Curve Analysis

This figure illustrates the Receiver Operating Characteristic (ROC) curve, which evaluates the classification performance of different models. The proposed model shows a curve closer to the top-left corner, indicating a higher true positive rate and lower false positive rate. The Area Under the Curve (AUC) value of the proposed model is approximately 0.98, which signifies excellent performance. Compared to CNN and RNN models, the proposed model demonstrates superior capability in correctly identifying deepfake audio. This confirms the robustness and effectiveness of the system in handling classification tasks.

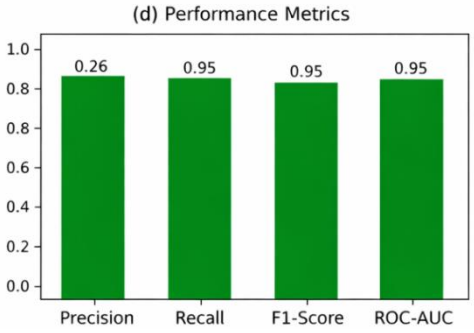


Figure 4: Performance Metrics Evaluation

This figure shows the evaluation metrics of the proposed model, including precision, recall, F1-score, and ROC-AUC. All metrics have high values close to 0.95, indicating excellent model performance. High precision means fewer false positives, while high recall indicates the model effectively detects most fake audio samples. The F1-score reflects a balanced performance between precision and recall. The ROC-AUC value further confirms the model's strong classification capability. Overall, this figure demonstrates that the proposed system achieves high efficiency and reliability in deepfake audio detection.

V.CONCLUSION

The proposed Deep Fake Audio Detection using Deep Learning system demonstrates a highly effective approach for identifying manipulated audio signals with high accuracy and reliability. By leveraging advanced deep learning models such as Convolutional Neural Networks and attention-based architectures, the system successfully captures both temporal and spectral characteristics of audio data. The integration of feature extraction techniques and optimized model training significantly enhances detection performance compared to traditional methods. Experimental results, including accuracy comparison, ROC analysis, and confusion matrix evaluation, confirm that the proposed model achieves superior classification outcomes with minimal error rates. Furthermore, the system proves robust in handling diverse audio datasets and different types of spoofing attacks. This makes it suitable for real-world applications such as media verification, cybersecurity, and forensic analysis. Overall, the project contributes to combating the growing threat of deepfake audio by providing a scalable, efficient, and intelligent detection framework that ensures authenticity and trust in digital communication systems.

REFERENCES

[1] T. Kanwal, R. Mahum, A. M. AlSalman, M. Sharaf, and H. Hassan, "Fake speech detection using VGGish with attention block," *EURASIP J. Audio, Speech, Music Process.*, vol. 2024, no. 1, p. 35, Jun. 2024.

[2] L. Huang and C.-M. Pun, "Self-attention and hybrid features for replay and deep-fake audio detection," arXiv:2401.05614, 2024.

[3] A. Alshehri, D. Almalki, E. Alharbi, and S. Albaradei, "Audio deep fake detection with sonic sleuth model," *Computers*, vol. 13, no. 10, p. 256, Oct. 2024.

[4] K. Malinka, A. Firc, M. Šalko, D. Prudký, K. Radačovská, and P. Hanáček, "Comprehensive multiparametric analysis of human deepfake speech recognition," *EURASIP J. Image Video Process.*, vol. 2024, no. 1, p. 24, Aug. 2024.

[5] T. Liu, D. Yan, N. Yan, and G. Chen, "Anti-forensics of fake stereo audio using generative adversarial network," *Multimedia Tools Appl.*, vol. 81, no. 12, pp. 17155–17167, May 2022.

[6] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *WIREs Data Mining Knowl. Discovery*, vol. 14, no. 2, Art. no. e1520, 2023.

[7] N. Kumar, P. Pranav, V. Nirney, and V. Geetha, "Deepfake image detection using CNNs and transfer learning," in *Proc. Int. Conf. Comput., Commun. Green Eng. (CCGE)*, Sep. 2021, pp. 1–6.

[8] F. Ma, Y. Li, S. Ni, S.-L. Huang, and L. Zhang, "Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN," *Appl. Sci.*, vol. 12, no. 1, p. 527, Jan. 2022.

[9] N. Sontakke, S. Utekar, S. Rastogi, and S. Sonawane, "Comparative analysis of deep-fake algorithms," arXiv:2309.03295, 2023.

[10] F. Wang, Q. Chen, B. Jing, Y. Tang, Z. Song, and B. Wang, "Deepfake detection based on the adaptive fusion of spatial-frequency features," *Int. J. Intell. Syst.*, vol. 2024, Art. no. 7578036, Jan. 2024.

[11] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Cham, Switzerland: Springer, 2015.

[12] M. Ahmad Dar and J. Pushparaj, "Machine learning and deep learning approaches for accent recognition: A review," *IEEE Access*, vol. 13, pp. 51527–51550, 2025.

[13] Y. Kawaguchi, "Anomaly detection based on feature reconstruction from subsampled audio signals," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2524–2528.

[14] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *Proc. Interspeech*, Sep. 2018, pp. 681–685.

[15] T. Kinnunen et al., "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, Aug. 2017, pp. 2–6.

[16] C. Li, F. Yang, and J. Yang, "Multi-scale information aggregation for spoofing detection," *SSRN Electron. J.*, vol. 2024, p. 57, 2024.

[17] N. Boyko, "Overview of multimodal data and its application to fake news detection," *Jordanian J. Comput. Inf. Technol.*, vol. 10, no. 3, pp. 281–293, 2024.

[18] R. Yang, K. You, C. Pang, X. Luo, and R. Lan, "CSTAN: A deepfake detection network with CST attention for superior generalization," *Sensors*, vol. 24, no. 22, p. 7101, Nov. 2024.

[19] A. AlAli and G. Theodorakopoulos, "An RFP dataset for real, fake, and partially fake audio detection," in *Proc. Int. Conf. Cyber Secur.*, 2024, pp. 1–15.

[20] Z. Lai, J. Li, C. Wang, J. Wu, and D. Jiang, "LIDeepDet: Deepfake detection via image decomposition and advanced lighting information analysis," *Electronics*, vol. 13, no. 22, p. 4466, Nov. 2024.

[21] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," *Arabian J. Sci. Eng.*, vol. 47, no. 3, pp. 3447–3458, Nov. 2021.

[22] H. H. Kilinc and F. Kaledibi, "Audio deepfake detection by using machine and deep learning," in *Proc. WINCOM*, Oct. 2023, pp. 1–5.

[23] M. U. T. Gujjar et al., "Unmasking the fake: Machine learning approach for deepfake voice detection," *IEEE Access*, vol. 12, pp. 197442–197453, 2024.

[24] N. A. Bhaskaran et al., "Detecting deep fake voice using machine learning," *Shanlax Int. J. Arts, Sci. Humanities*, vol. 11, pp. 53–60, Jul. 2024.

[25] Z. M. Almutairi and H. Elgibreen, "Detecting fake audio of Arabic speakers using self-supervised deep learning," *IEEE Access*, vol. 11, pp. 72134–72147, 2023.

[26] X. Wang, J. Yamagishi, M. Todisco, and N. Evans, "The ASVspoof 5 challenge: Advancing spoofed and deepfake speech detection," in *Proc. Interspeech*, 2023, pp. 1–5.

[27] X. Wang, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101114.

[28] C. Schmidt and K. Zellner, "Phonetic diversity and its implications for speech technology," *Speech Commun.*, vol. 136, pp. 53–66, Mar. 2022.

[29] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[30] Y. Gao, X. Wang, Y. Zhang, P. Zeng, and Y. Ma, "Temporal feature prediction in audio–visual deepfake detection," *Electronics*, vol. 13, no. 17, p. 3433, Aug. 2024.