International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

VocaTrust: Machine Learning Based Voice Integrity Assessment

Prof.Yashodara R[1], Deeksha[2], Dimple G[3] ,Sinchana Gowda RP[4] , Varsha NG[5]

[1](Department of Information Science and Engineering, Don Bosco Institute of Technology, Bengaluru)
[2,3,4,5](Department of Information Science and Engineering, Don Bosco Institute of Technology, Bengaluru)

----------**********----------

Abstract:

The rapid growth of artificial intelligence has made it possible to generate highly realistic synthetic voices that closely resemble human speech. While this technology has many useful applications, it has also increased the risk of voice-based fraud, identity theft, misinformation, and unauthorized access. These challenges highlight the need for accurate and efficient deepfake voice detection systems.

This paper presents VocaTrust: Machine Learning-Based Voice Integrity Assessment, a framework designed to detect manipulated speech and evaluate voice authenticity. The system uses genuine audio samples from the LibriSpeech dataset and spoofed audio from the ASVspoof dataset. After preprocessing, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from each audio sample to capture key speech characteristics. These features are then used to train a Convolutional Neural Network (CNN) that classifies speech as genuine or fake. In addition, the model generates a Voice Integrity Score that indicates the likelihood of the audio being authentic.

The proposed framework is lightweight, computationally efficient, and suitable for real-time applications using only audio input. By providing reliable deepfake voice detection with low computational overhead, VocaTrust can support secure voice authentication, banking systems, digital forensics, and voice-enabled applications against emerging spoofing attacks.

Keywords - **Deepfake Voice Detection, Voice Integrity Assessment, Machine Learning, CNN, MFCC, Speech Spoofing.**

----------**********----------

I. INTRODUCTION

Artificial Intelligence (AI) has significantly advanced speech synthesis, enabling the creation of highly realistic synthetic voices through modern deep learning techniques. Recent text-to-speech (TTS) and voice cloning models can accurately imitate a person's tone, pitch, pronunciation, and speaking style, making AI-generated speech increasingly difficult to distinguish from genuine human voices. While these technologies have improved applications such as virtual assistants, accessibility tools, entertainment, and digital communication, they have also introduced serious security and ethical concerns. Synthetic voices are now being misused for identity impersonation, financial fraud, misinformation, social engineering, and unauthorized access to voice-based authentication systems, highlighting the need for reliable methods to verify voice authenticity.

The rapid rise of deepfake audio has become a major concern in cybersecurity and digital forensics. Conventional speaker verification systems primarily focus on recognizing a speaker's identity and often fail to detect AI-generated or manipulated speech.

Although several deepfake detection methods have been proposed, many perform only binary classification, require significant computational resources, rely on multimodal inputs such as audio and video, or show limited performance when processing noisy recordings and previously unseen spoofing attacks. These limitations reduce their effectiveness in practical environments where accurate, efficient, and real-time detection is essential.

To address these challenges, this paper proposes VocaTrust: Machine Learning-Based Voice Integrity Assessment, a lightweight machine learning framework for detecting manipulated speech and assessing voice authenticity. The system uses genuine speech from the LibriSpeech dataset and spoofed speech from the ASVspoof dataset. After preprocessing, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to represent the acoustic characteristics of the audio, and these features are used to train a Convolutional Neural Network (CNN) for classifying speech as genuine or fake. In addition to classification, the proposed framework generates a Voice Integrity Score, providing a confidence-based measure of authenticity. Designed as an audio-only solution, VocaTrust is computationally efficient, suitable for real-time deployment, and applicable to voice authentication, banking security, digital forensics, call verification, and the protection of voice-enabled systems.

II.MATERIALS AND METHODS

The proposed VocaTrust system was developed using publicly available speech datasets containing both genuine and AI-generated audio samples. Genuine speech recordings were obtained from the LibriSpeech corpus, whereas synthetic speech samples were sourced from the ASVspoof dataset, which includes speech generated using various Text-to-Speech (TTS) and Voice Conversion (VC) techniques. The proposed framework was implemented in Python using the TensorFlow/Keras deep learning framework. Audio preprocessing and feature extraction were carried out using the Librosa library, while NumPy, Pandas, and Scikit-learn were used for data processing, model evaluation, and performance analysis. All experiments were performed on Google Colaboratory, utilizing an NVIDIA Tesla T4 GPU to accelerate model training and testing.

The proposed VocaTrust system follows a structured methodology consisting of multiple stages, including data collection, preprocessing, feature extraction, model design, training, and voice integrity scoring. Initially, genuine and synthetic audio samples are collected from publicly available datasets such as LibriSpeech and ASVspoof. These datasets provide a diverse collection of real and AI-generated speech recordings, enabling effective model training and evaluation. To improve the model's robustness under real-world acoustic conditions.

In the preprocessing stage, all audio files are converted into a standardized format to ensure consistency across the dataset. The preprocessing operations include mono conversion, resampling to 16 kHz, normalization of audio duration, and conversion into WAV format. These steps help minimize variations caused by different recording conditions and improve the reliability of feature extraction. After preprocessing, Mel Frequency Cepstral Coefficients (MFCCs) are extracted from the audio signals. MFCC features effectively capture important speech characteristics, such as pitch, tone, and frequency distribution, while reducing redundant information. Their ability to provide compact and meaningful representations of speech signals makes them well suited for deepfake speech detection.

The extracted MFCC features are then provided as input to a Convolutional Neural Network (CNN) model. The CNN architecture is designed to learn complex speech patterns and identify subtle differences between genuine and synthetic audio recordings. The model is trained using supervised learning techniques with labeled real and fake audio samples. During training, the network optimizes its parameters using the Adam optimizer and the Binary Cross-Entropy loss function. To enhance the transparency of the classification process, Score-CAM is incorporated as an Explainable Artificial

Intelligence (XAI) technique to visualize the regions of the input features that contribute most to the model's prediction.

Finally, the trained model generates a Voice Integrity Score representing the confidence level of speech authenticity. Based on this score, the system classifies each audio sample into categories such as genuine, suspicious, or fake. The integration of robust preprocessing, CNN-based classification, Explainable AI, and voice integrity scoring enables the proposed methodology to provide a lightweight, interpretable, and real-time deepfake audio detection system suitable for practical deployment in real-world applications.

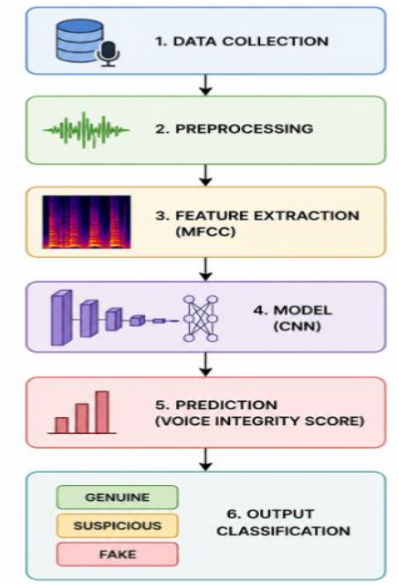

fig. 1. VocaTrust Methodology Workflow

III. LITERATURE SURVEY

Several researchers have proposed different machine learning and deep learning techniques for detecting.AI-generated speech and audio deepfakes. These studies mainly focus on improving detection accuracy, robustness, and generalization using various feature extraction methods and neural network architectures.

The following literature review summarizes some of the significant contributions related to deepfake voice detection.

1. Naveed, S. J., et al. (2026) [1] This study proposed DEEPSHIELD, a multimodal deep learning framework that combines audio and video information using CNN, BiLSTM, GRU, and cross-modal attention mechanisms. The model achieved high detection accuracy by learning both facial and speech features but required both audio and video inputs, increasing computational complexity.

2. Jung, J. W., et al. (2025) [3] This research introduced the SpoofCeleb dataset for speech deepfake detection using large-scale real andsynthetic voice recordings. The study evaluated several deep learning models and demonstrated improved robustness for detecting spoofed speech in real-world environments.

3. Schäfer, K., and Steinebach, M. (2025) [4] The authors evaluated Automatic Speech Recognition (ASR) encoder-based feature extraction techniques using Whisper, SpeechT5, Wav2Vec2, and CNN models. Their results showed that deep learning-based acoustic representations improved deepfake detection performance compared to traditional handcrafted features.

4. Guragain, A., et al. (2024)[8] This work proposed an ensemble framework using self-supervised learning models and AASIST for singing voice deepfake detection. The combination of multiple speech representations improved detection accuracy and reduced error rates across different evaluation datasets.

5. Zhang, Y., et al. (2024) [9] The authors introduced the SVDD2024 challenge and benchmark datasets for singing voice deepfake detection. The study provided a standardized evaluation framework for comparing different deep learning models under realistic recording conditions.

6. Chaiwongyen, A., et al. (2023) [12] This research presented a deepfake speech detection method using pathological speech features such as jitter, shimmer, and harmonic-to-noise ratio combined with a Multilayer Perceptron (MLP) classifier. The proposed approach achieved good detection accuracy while remaining computationally efficient.

7. Yadav, A. K. S., et al. (2023) [14]The proposed PS3DT framework utilized Mel-spectrogram patches and Transformer networks for synthetic speech detection. The model demonstrated improved generalization and robust performance against different speech synthesis techniques.
8. Yan, R., et al. (2022) [15] This study developed an audio deepfake detection system using ResNet-34, multi-head attention, neural stitching, and LFCC features. The proposed framework enhanced detection accuracy and improved robustness against unseen spoofing attacks.

IV.EXPECTED CONCLUSION

A.Detection Performance:
The proposed VocaTrust system is expected to accurately distinguish between genuine and AI-generated speech by utilizing Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction and a Convolutional Neural Network (CNN) for classification. By learning the subtle acoustic differences between real and synthetic speech, the model is anticipated to achieve reliable deepfake voice detection with improved accuracy. This capability is expected to enhance the trustworthiness of automatic voice verification systems and contribute to more dependable speech authenticity assessment.

B. Voice Integrity Assessment:
The proposed system generates a Voice Integrity Score ranging from 0 to 100, representing the confidence level of the authenticity of an input voice sample. Unlike conventional approaches that provide only a binary real-or-fake decision, the proposed scoring mechanism offers a confidence-based assessment, enabling users to better understand the reliability of the prediction. This additional level of interpretation makes the system more transparent and practical for real-world decision-making.

C. Practical Applications:
The proposed framework is expected to be applicable in a wide range of real-world domains, including voice authentication, banking security, digital forensics, fraud detection, and secure communication systems. Owing to its lightweight CNN-based architecture and efficient processing pipeline, the system is suitable for practical deployment in environments where fast and reliable voice verification is essential. Furthermore, the integration of explainable AI enhances user confidence by providing greater transparency into the model's predictions.

D. Final Assessment:
Overall, the proposed VocaTrust system is expected to provide an efficient, reliable, and interpretable solution for deepfake voice detection. By combining robust preprocessing, MFCC-based feature extraction, CNN-based classification, Explainable AI, and Voice Integrity Scoring, the framework offers a comprehensive approach for assessing speech authenticity. The proposed methodology also provides opportunities for future enhancements, such as incorporating multilingual support, expanding the system to handle more diverse speech datasets, and enabling real-time voice analysis for practical applications in dynamic environments.

ACKNOWLEDGMENT

This work was supported by the faculty of Information Science and Engineering, Don Bosco Institute of Technology, Bengaluru. The authors would like to express their gratitude to ASVspoof dataset.

REFERENCES

[1] Naveed, S. J., Alif, A. I., & Saha, S. (2026). DEEPSHIELD: A Multimodal Deep Learning Model for Audio-Visual Deepfake Detection. *Proceedings of the 5th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, pp. 1–6.

[2] Wani, T. M., & Amerini, I. (2026). Multi-Scale Self-Supervised Learning for Efficient Audio Deepfake Detection. *IEEE Signal Processing Letters*, 33, pp. 46–50.

[3] Jung, J. W., et al. (2025). SpoofCeleb: Speech Deepfake Detection and SASV in the Wild. *IEEE Open Journal of Signal Processing*, 6, pp. 68–77.

[4]Schäfer, K., & Steinebach, M. (2025). Generalization in Audio Deepfake Detection: Evaluating ASR Encoder-Based Feature Extraction. *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–5.

[5] Coletta, E., Salvi, D., Negroni, V., Leonzio, D. U., & Bestagini, P. (2025). Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching. *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 820–824.

[6] Pham, L., Tran, D., Lam, P., Skopik, F., Schindler, A., Poletti, S., Fischinger, D., & Boyer, M. (2025). DIN-CTS: Low-Complexity Depthwise-Inception Neural Network with Contrastive Training Strategy for Deepfake Speech Detection. *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 556–560.

[7] Chauhan, S., Sethi, K., & Anand, A. (2025). Audio Deepfakes Detection: A Comprehensive Literature Review of Voice Cloning Fraud Detection Techniques. *Proceedings of the IEEE 3rd International Symposium on Sustainable Energy, Signal Processing and Cybersecurity (iSSSC)*, pp. 1–5.

[8] Guragain, A., Liu, T., Pan, Z., Sailor, H. B., & Wang, Q. (2024). Speech Foundation Model Ensembles for the Controlled Singing Voice Deepfake Detection (CTRSVDD) Challenge 2024. *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pp. 774–781.

[9] Zhang, Y., Zang, Y., Shi, J., Yamamoto, R., Toda, T., & Duan, Z. (2024). SVDD 2024: The Inaugural Singing Voice Deepfake Detection Challenge. *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pp. 782–787.

[10] Chiddarwar, P. (2024). Real-Time Detection of AI-Generated Deepfake Audio: A Novel Approach. *Proceedings of the IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, pp. 1–5.

[11] Jain, E., & Singh, A. (2024). Deepfake Voice Detection Using Convolutional Neural Networks: A Comprehensive Approach to Identifying Synthetic Audio. *Proceedings of the 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*, pp. 1–5.

[12] Chaiwongyen, A., Duangpummet, S., Karnjana, J., Kongprawechnon, W., & Unoki, M. (2023). Deepfake Speech Detection with Pathological Features and Multilayer Perceptron Neural Network**.** *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2182–2188.

[13] Salvi, D., Bestagini, P., & Tubaro, S. (2023). Reliability Estimation for Synthetic Speech Detection. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5.

[14] Singh Yadav, A. K., Xiang, Z., Bhagtani, K., Bestagini, P., Tubaro, S., & Delp, E. J. (2023). PS3DT: Synthetic Speech Detection Using Patched Spectrogram Transformer**.** *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pp. 496–503.

[15] Yan, R., Wen, C., Zhou, S., Guo, T., Zou, W., & Li, X. (2022). Audio Deepfake Detection System with Neural Stitching for ADD 2022. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9226–9230.

[16] Cui, S., Huang, B., Huang, J., & Kang, X. (2022). Synthetic Speech Detection Based on

Local Autoregression and Variance Statistics**.** *IEEE Signal Processing Letters*, **29**, pp. 1462–1466.

[17] Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M. C., & Tubaro, S. (2022). Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach**.** *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8962–8966.

