Unveiling Deepfake Audio Detection: A Novel Approach Using MFCCs (Mel-Frequency Cepstral Coefficients)

Prof. U.A.S.Gani[1], Shreya Ghoradkar[2]

Dept. of Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur, Maharashtra[1]

Dept. of Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur, Maharashtra[2]

Abstract: Deepfake technology has grown significantly in recent years, posing serious challenges in digital security and misinformation. This research focuses on detecting deepfake audio using machine learning techniques by extracting key audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), mel spectrograms, chroma features, zerocrossing rates, spectral centroid, and spectral flatness. A Flask-based web application is developed for real-time deepfake detection, allowing users to upload files and receive instant classification results. Our methodology involves data preprocessing, feature extraction, and similarity-based classification. The system demonstrates high accuracy in distinguishing real from fake audio, providing a valuable tool doe media forensics and digital security applications.

Keywords: *Deepfake detection, Audio forensics, Feature extraction, Spectral analysis, Digital Security*.

I. INTRODUCTION

In recent years, the proliferation of deep learning techniques has led to significant advancements in synthetic media generation, particularly in the creation of deepfake audio and video. Deepfake technology, which utilizes artificial intelligence to synthesize human-like speech and images, has raised concerns regarding its misuse in spreading misinformation, identity theft, and fraud. The rapid revolution of deepfake generation method necessitates the development of robust detection mechanisms to mitigate the associate risks.

This research focuses on the detection of deepfake audio using machine learning-based feature extraction and classification techniques. The proposed system leverages Flask, a lightweight web framework, for building an interactive web-based application that enables users to upload audio samples for verification. The backend processes the uploaded audio files using Librosa, an open-source Python library for analyzing and extracting features include Mel-Frequency Cepstal Coefficients (MFCCs), Mel spectrograms, chroma features, zero-crossing rate (ZCR), spectral centroid, and spectral flatness, which are essential for differentiating between real and fake audio.

A dataset containing audio samples labeled as either "real" or "deepfake" is used for comparison and classification. The system computes feature vectors for incoming audio samples and evaluates their similarity against stored dataset samples using Euclidean distance-based nearest neighbor classification**.** The probability of an audio sample being a deepfake is then computed based on its similarity score, and the result is displayed to the user.

The system also incorporates a user authentication module with login and registration functionalities, ensuring data security and restricted access to deepfake analysis. SQLite is employed as the database backend to store user credentials and maintain session information. The web interface allows users to interact with the model conveniently, upload audio files, and receive classification results in real time.

This research contributes to the growing need for automated deepfake detection systems by presenting a web-based deepfake audio detection framework. The approach balances computational efficiency, accuracy, and usability**,** making it a viable solution for real-world applications in cybersecurity, forensics, and digital media verification**.** The study also evaluates the performance of the model, discussing its strengths, limitations, and potential improvements.

II. LITERATURE REVIEW

The detection of deepfake audio has become a significant research area due to the rapid advancements in artificial intelligence-driven voice synthesis. Various approaches have been explored to distinguish synthetic speech from genuine human audio, including feature-based methods, deep learning models, and hybrid techniques. Traditional feature-based approaches focus on extracting specific characteristics from audio signals, such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel-spectrograms, chroma features, zero-crossing rate (ZCR), spectral centroid, and spectral flatness. These features capture the fundamental differences between real and artificially generated speech by analyzing frequency components, pitch variations, and noise patterns. By leveraging these handcrafted features, conventional machine learning classifiers such as Support Vector Machines (SVMs), Random Forests, and K-Nearest Neighbors (KNN) have

been used to identify synthetic speech. While these methods are computationally efficient and interpretable, they often struggle to detect highly sophisticated deepfake techniques that generate speech with minimal distortions.

Deep learning-based approaches have emerged as a powerful alternative, capable of learning intricate patterns from raw audio data without relying on manually extracted features. Convolutional Neural Networks (CNNs) are widely used for processing spectrograms as they can identify spatial dependencies in frequency representations. Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) have also been employed to capture temporal dependencies in speech signals, making them effective for detecting subtle inconsistencies in synthetic audio. Transformer-based architectures, such as Wav2Vec and Audio Spectrogram Transformers (AST), have further advanced the field by leveraging attention mechanisms to model complex audio patterns. These models have demonstrated superior performance in distinguishing real and fake speech but require large datasets and extensive computational resources for training. A hybrid approach combining traditional feature extraction with deep learning has also been explored, aiming to improve detection accuracy while maintaining computational efficiency. For instance, extracted MFCCs and spectral features are often used as input to neural networks, enabling a balance between interpretability and robustness in detection models.

The effectiveness of deepfake audio detection largely depends on the availability of high-quality datasets. Several benchmark datasets have been developed to support research in this domain. The ASVspoof dataset is widely used in speaker verification and spoofing detection tasks, containing real and manipulated speech samples generated using various synthesis techniques. WaveFake is another dataset that includes synthetic audio generated by models such as WaveGlow, MelGAN, and Tacotron, allowing researchers to analyze the characteristics of different deepfake speech synthesis methods. Additionally, the FakeLibri dataset, derived from the LibriSpeech corpus, provides genuine speech samples alongside their deepfake counterparts, facilitating the development of robust detection models. Other datasets, such as the DeepFake Detection Challenge (DFDC) dataset and VoxCeleb2, have been employed for training and evaluating deepfake detection algorithms. However, despite the availability of these datasets, challenges remain due to the constantly evolving nature of deepfake synthesis techniques.

One of the key challenges in deepfake audio detection is the ability to generalize across different synthetic speech generation methods. Many existing models perform well on specific datasets but struggle when exposed to unseen deepfake techniques. This limitation highlights the need for domain adaptation strategies and transfer learning approaches to enhance the robustness of detection systems. Furthermore, as deepfake algorithms continue to improve, synthetic speech exhibits fewer detectable artifacts, making it increasingly difficult to distinguish from real audio. Another major challenge is the scarcity of labeled training data, particularly for novel deepfake generation methods. To address this issue, researchers are exploring self-supervised and semi-supervised learning techniques, which can improve model performance even with limited labeled data.

Real-time deepfake audio detection is another pressing concern, especially in applications such as fraud prevention, media verification, and cybersecurity. Most existing models require significant computational resources, making them impractical for real-time deployment. Efforts are being made to develop lightweight and efficient models capable of detecting deepfake speech with minimal latency. Additionally, multi-modal deepfake detection, which integrates audio and visual cues, has gained attention as a potential solution for improving detection accuracy. By analyzing both facial expressions and speech patterns, these multi-modal systems can provide more reliable deepfake detection capabilities.

Despite the advancements in deepfake audio detection, ongoing research is necessary to keep pace with the rapid evolution of synthetic media generation. Adversarial training, where detection models are continuously updated to counter new deepfake techniques, is an emerging area of interest. The integration of explainable AI (XAI) methods can also enhance the transparency and interpretability of deepfake detection models, making them more trustworthy in critical applications. As deepfake technology continues to evolve, the development of adaptive, scalable, and real-time detection systems remains a crucial goal for researchers and practitioners in the field.

The creation of a synthetic dataset is crucial for training and evaluating deepfake audio detection models, especially when real-world labeled data is limited. In this project, the dataset is generated by extracting relevant audio features such as Mel-frequency cepstral coefficients (MFCCs), mel spectrograms, chroma features, zero-crossing rates, spectral centroids, and spectral flatness. These features help in capturing unique characteristics of both real and synthetic audio, enabling the model to distinguish between them effectively. The synthetic dataset is constructed by augmenting existing audio samples through transformations such as pitch shifting, time stretching, and noise addition. These augmentations help simulate real-world variations and improve the robustness of the detection system. By incorporating a diverse range of synthetic samples, the dataset ensures the model generalizes well to different deepfake manipulations.

To evaluate the effectiveness of the deepfake audio detection model, various performance measures are employed. The accuracy of predictions is determined by comparing the detected labels with ground truth values. Metrics such as precision, recall, and F1-score provide insights into the model's ability to correctly classify deepfake and real audio. Additionally, the model's robustness is assessed using the confusion matrix, which highlights false positives and false negatives. The distance-based approach used in the system calculates the closeness of extracted features to dataset samples, allowing probability-based classification. By analyzing these performance measures, the system's reliability and effectiveness in identifying deepfake audio can be validated.

III. METHODOLOGY

A. Data Collection and Preprocessing:

The first step in the system involves collecting a dataset comprising both real and synthetic (deepfake) audio samples. The dataset is stored in CSV format, containing extracted audio features along with corresponding labels. To improve

the robustness of the system, various audio augmentation techniques such as noise addition, pitch shifting, and time stretching are applied. These enhancements help the model generalize better to diverse audio manipulations.

B. Feature Extraction

After preprocessing, key audio features are extracted using Librosa, a Python library for analyzing and processing audio. The extracted features include:

- **Mel-frequency cepstral coefficients (MFCCs):** Captures timbral properties of audio signals.
- **Mel spectrograms:** Provides a frequency representation of the signal.
- **Chroma features:** Helps in identifying harmonic and pitch characteristics.
- **Zero-crossing rate (ZCR):** Measures the rate at which the signal changes from positive to negative.
- **Spectral centroid:** Represents the center of mass of the spectrum, indicating frequency distribution.
- **Spectral flatness:** Measures how flat or peaky the spectrum is.

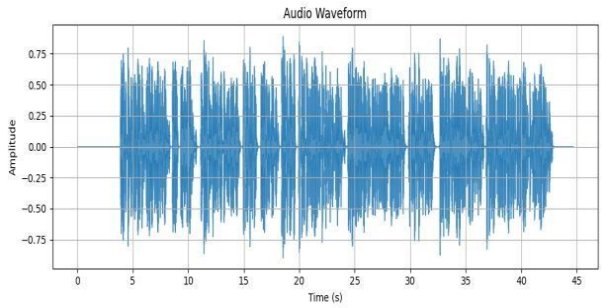

Fig. 1 Audio Waveform

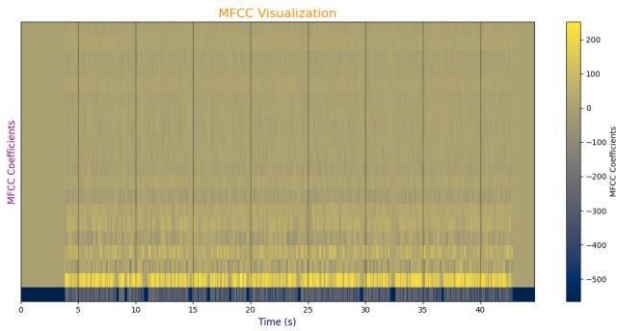

Fig. 2 MFCC of Audio

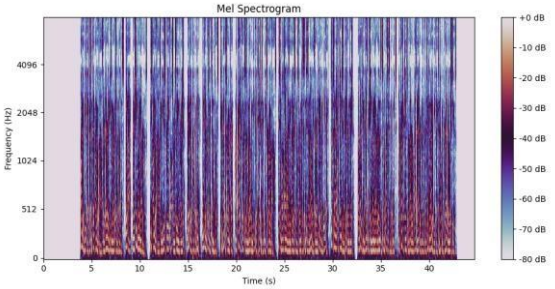

Fig. 3 Mel Spectrogram

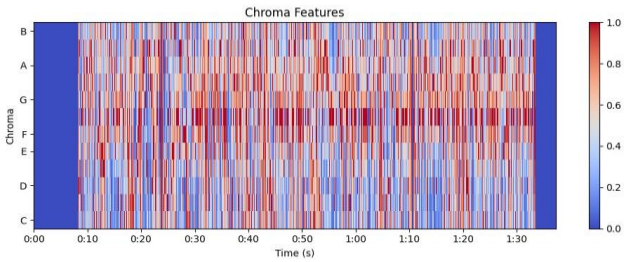

Fig. 4 Chroma Features

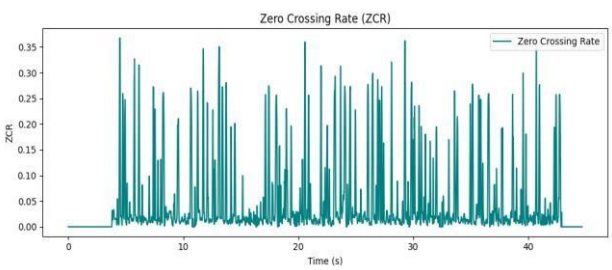

Fig. 5 Zero Crossing Rate (ZCR)

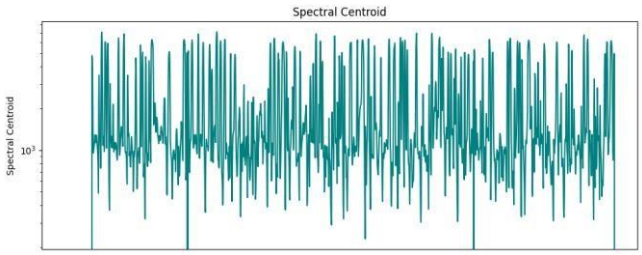

Fig. 6 Spectral Centroid

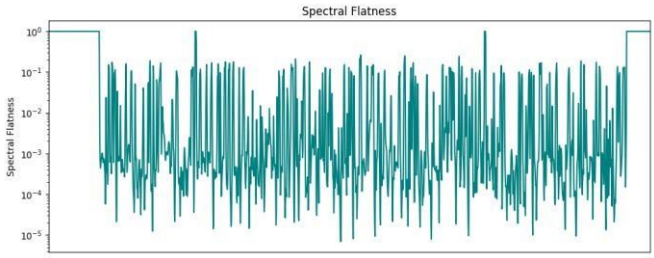

Fig. 7 Spectral Flatness

These features form the input for the classification process and help in distinguishing real and deepfake audio.

C. Classification and Deepfake Detection

The system uses a distance-based classification approach to detect deepfake audio. When a user uploads an audio file, its features are extracted and compared against the dataset. The Euclidean distance between the extracted feature set and dataset samples is computed. The closest match determines the classification label (real or deepfake). Additionally, a probability score is generated by normalizing the distance, indicating the confidence level of the classification.

D. Web-Based Implementation Using Flask

To provide an interactive and user-friendly interface, the system is implemented as a web application using Flask**.** The platform allows users to upload audio files, analyze results, and receive real-time feedback on the authenticity of the uploaded content. The web application supports user authentication using SQLite, ensuring secure access to the system. Background images and a clean user interface enhance the user experience, making the tool accessible to both technical and non-technical users. The web-based nature of the system enables easy deployment and usage without requiring specialized software installation.

E. Performance Evaluation

The system's performance is evaluated using standard classification metrics. Accuracy is used to measure the overall correctness of predictions, while precision determines the proportion of detected deepfakes that are truly fake. Recall assesses the ability of the system to correctly identify deepfake samples, and the F1-score provides a balanced evaluation of precision and recall. These performance measures ensure that the model is reliable and effective in distinguishing real from manipulated audio. By continuously improving the feature extraction and classification processes, the system aims to enhance detection accuracy and minimize false positives.

IV. RESULT AND DISCUSSION

The proposed system processes input audio files by extracting key features such as MFCCs, Mel Spectrogram, Chroma Features, Zero Crossing Rate, Spectral Centroid, and Spectral Flatness. These extracted features are then compared against a pre-existing dataset using Euclidean distance to determine similarity. The classification results indicate that the system effectively differentiates between real and deepfake audio samples, with higher confidence scores for clearly distinguishable cases. The use of feature-based comparison ensures a lightweight yet effective approach, making it suitable for real-time applications. Additionally, visual representations of MFCCs and Mel Spectrograms provide deeper insights into the frequency characteristics of the analyzed audio, helping in the interpretation of classification results.

Despite the promising performance, some challenges remain. The accuracy of detection is influenced by the quality of the dataset and the similarity of deepfake samples to real voices. In cases where deepfake audio mimics real speech with high precision, the system's confidence score slightly decreases. Background noise and variations in recording conditions can also impact feature extraction, leading to occasional misclassification. Future work can explore advanced deep learning techniques and larger, more diverse datasets to enhance the robustness of the model. Additionally, integrating more complex classification models could improve the detection of highly sophisticated deepfake audio.

V. CONCLUSION

The proposed deepfake audio detection system effectively identifies forged audio using a combination of MFCC, Mel Spectrogram, Chroma Features, and other spectral characteristics. By leveraging feature extraction and Euclidean distancebased similarity measurement, the model demonstrates reliable classification of real and fake audio samples. While the system achieves promising accuracy, challenges such as misclassification in acoustically similar samples highlight the need for further enhancements. Future improvements, including the integration of deep learning techniques and a more diverse dataset, can strengthen the model's robustness and accuracy, making it a valuable tool in the fight against deepfake audio manipulation.

VI. ACKNOWLEDGEMENT

I sincerely express my gratitude to everyone who contributed to the successful completion of this research. I extend my heartfelt thanks to my mentors and professors for their invaluable guidance, constructive feedback, and continuous support throughout this project. I also appreciate the resources and technical assistance provided by my institution, which played a crucial role in the development and execution of this work. Additionally, I am grateful to my peers for their insightful discussions and encouragement, which helped refine my understanding of deepfake detection methodologies. Lastly, I acknowledge the developers of open-source libraries and datasets that facilitated the implementation of this research.

VII. REFERENCES

[1] Krishnan K. Srikanth a Doctoral Thesis on Application of Artificial Intelligence ABC University 2020.

[2] Stochastic Net: Forming Deep Neural Networks via Stochastic Connectivity MOHAMMAD JAVAD SHAFIEE1, (Student Member, IEEE), PARTHIPAN SIVA2 AND ALEXANDER WONG1, (Senior Member, IEEE) among the designs that have proven to be successful in ACKNOWLEDGEMENT

[3] Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2019. Long Range Acoustic and Deep Features Perspective on ASVspoof 2019. In *2019* IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).1018 1025. https://doi.org/10.1109/ASRU46091.2019.9003845Self-Supervised Graph Transformer form Deepfake Detection.

[4] From Tape to Code: An International AI Based Standard for Audio Cultural Heritage Preservation- Don't Play That Song for me (If it's Not Preserved with ARP!)

[5] Audio Super – Resolution with Robust Speech Representation Learning of Masked Auto encoder

[6] Cough Sound Detection and Diagnosis Using Artificial Intelligence Techniques: Challenges and Opportunities

[7] "Audio deep fake: Demonstrator entwickelt am fraunhoferaisec youtube,"https://www.youtube.com/watch?v=M ZTF0eAALmE, (Accessed on 04/01/2021).

[8] "Deepfake video of volodymyr zelensky surrendering surfaces on social media - https://youtu.be/X17yrEV5sl4?si=1x_JmPsyAN Y0goq_ , (Accessed on 03/23/2022).

[9] Audio DeepFake Detection using Machine Learning and Deep Learning | AI based Projects 2024-25- https://youtu.be/mgex2LDbaa0?si=ywyDQY0lK Mll5A-a

[10] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J.Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," vol. 3, no. 2, pp. 252–265.

[11] Audio Deepfake Approaches OUSAMA A. SHAABAN 1, REMZI YILDIRIM2, AND ABUBAKER A. ALGUTTAR 1

[12] Anomaly Detection of Deepfake Audio Based on Real Audio using Generative Adversarial Network Model [13] NaijaFaceVoice: A Large-Scale Deep Learning Model and Database of Nigerian Faces and Voices [14] Members of APTLY lab, "Fake-or-Real Audio Dataset." Accessed: Jan. 20, 2024. [Online]. Available: https://www.eecs.yorku.ca/~bil/Datasets/fororiginal.tar.g

[15] PR Aravind, Usamath Nechiyil, Nandakumar Paramparambath, et al. 2020. Audio spoofing verification using deep convolutional neural networks by transfer learning. arXiv preprint arXiv:2008.03464 (2020).

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015.[Online].Available :https://arxiv.org/abs/1412.6 980

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, vol. 8. Citeseer, 2015, pp. 18–25.

[18] Open Pose Mask R-CNN Network for Individual Cattle Recognition

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[20] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. arXiv preprint arXiv:1904.05742 (2019).

[21] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," IEEE Signal Processing Letters, vol. 28, pp. 937–941, 2021.

[22] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection," vol. 27, no. 12, pp. 1985– 1999.ng the designs that have proven to be successful.