International Journal of Advance Research in Science and Engineering
Volume. No. 15, Issue No. 03, March, 2026
ISSN: 2319-8354

Deep Fake Audio Detection Using Deep Learning

Guide: Mr.K Vinod kumar, B.Tech, M.Tech, Associate Professor

Vankayalapati Pujitha
Department of IT
Tirumala Engineering College
Narsaraopet, 522601
vankayalapatipujitha08@gmail.com

Mandapalli Srihari
Department of IT
Tirumala Engineering College
Narsaraopet, 522601
mandapallisrihari9999@gmail.com

Katta Sravan Ajay Kumar
Department of IT
Tirumala Engineering College
Narsaraopet, 522601
ajaychowdarykatta110125@gmail.com

Yaragarla Venkata Kavya
Department of IT
Tirumala Engineering College
Narsaraopet, 522601
yaragarla17@gmail.com

(Academic Year: 2022 – 2026)

Abstract - The rapid advancement of artificial intelligence has enabled the creation of highly realistic synthetic audio, commonly known as audio deepfakes. These manipulated audio clips pose serious threats in areas such as misinformation, voice fraud, and digital security. This project presents an efficient and practical Audio Deepfake Detection System using machine learning techniques to distinguish between real and fake audio signals.

The system implements a dual-algorithm approach combining Support Vector Machine (SVM) with Mel-Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Model (GMM) with Constant-Q Cepstral Coefficients (CQCC). These feature extraction techniques capture unique spectral and frequency characteristics of audio signals, enabling accurate classification. A real-time web-based interface is developed using Streamlit, allowing users to upload audio files and receive instant detection results along with confidence scores and visual analysis.

Experimental results demonstrate high detection accuracy, achieving approximately 96.5% using SVM and 90.4% using GMM-based methods. The system provides a practical solution for real-world applications such as media verification, fraud detection, and digital forensics. This project highlights the importance of AI-based security systems in combating emerging threats from synthetic media.

I . Introduction

The rapid advancement of artificial intelligence and machine learning technologies has led to the emergence of highly sophisticated synthetic media, commonly known as deepfakes. Among these, audio deepfakes have gained significant attention due to their ability to mimic human voices with high accuracy. Using techniques such as voice cloning and neural speech synthesis, it is now possible to generate audio that sounds almost identical to real human speech. While these technologies have beneficial applications in areas like virtual assistants and entertainment, they also pose serious threats to security, privacy, and trust in digital communication.

Audio deepfakes can be misused for malicious purposes such as impersonation, spreading misinformation, financial fraud, and manipulation of public opinion. For example, attackers can generate fake voice recordings of individuals to deceive people or organizations. As a result, distinguishing between genuine and synthetic audio has become a critical challenge in today's digital world. Traditional methods of detecting fake audio rely on manual inspection or basic signal processing techniques, which are are often inefficient, time-consuming, and unable to handle modern deepfake technologies.

To address these challenges, there is a growing need for automated and intelligent systems capable of detecting audio deepfakes accurately and in real time. This project presents an Audio Deepfake Detection System using machine learning techniques that can effectively classify audio as real or fake. The system utilizes two powerful approaches:
Support Vector Machine (SVM) with Mel-Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Model (GMM) with Constant-Q Cepstral Coefficients (CQCC). These feature extraction methods capture the unique spectral and frequency characteristics of audio signals, enabling precise analysis and classification.

In addition to the detection models, a user-friendly web-based interface is developed using Streamlit, allowing users to upload audio files and obtain instant results along with confidence scores and visual representations. This enhances the usability of the system and makes it suitable for real-world applications. The proposed system not only improves detection accuracy but also provides a practical solution for combating the growing threat of audio deepfakes.

Overall, this project contributes to the field of artificial intelligence and cybersecurity by providing an efficient, reliable, and accessible tool for detecting synthetic audio, thereby helping to ensure authenticity and trust in digital media.

II. Literature Review

1.Tamilselvan G., Biswal M. (2025) – Voice Cloning & Deep Fake Audio Detection Using Deep Learning

This study provides a comprehensive overview of voice cloning and fake audio detection using deep learning techniques. It explains how neural network models like Tacotron and WaveNet generate realistic synthetic speech. The research highlights the importance of detecting manipulated audio using features such as spectrograms, pitch, and linguistic patterns, achieving high accuracy in identifying fake audio.

2.Abbasi A., Javed R. et al. (2022) – Audio Forensics Dataset for Anomaly Detection

This research introduces a large-scale dataset for anomaly detection in audio forensics. It supports the development of

machine learning models to detect rare and suspicious audio events. The study emphasizes that high-quality datasets are essential for improving the performance of deepfake detection systems.

3.Javed R., Ahmed W.et al. (2022) – Survey on Computer Forensics

This research presents a comprehensive survey of tools and techniques used in computer forensics. It highlights the challenges in detecting manipulated digital content and emphasizes the need for advanced machine learning approaches in audio and multimedia forensics.

4. A hmed S. et al. (2022) – Speaker Identification Using Deep Neural Networks

This study focuses on speaker recognition using deep neural networks. It highlights the role of voice features and deep learning models in identifying speakers, which is closely related to detecting fake.

5.Anwar S. et al. (2022) – Social Data Analysis Using Embeddings

This study explores advanced embeddings for analyzing complex data patterns. It indirectly supports deepfake detection by demonstrating how feature representations can improve classification accuracy in machine learning models.

6.Kawaguchi (2021) – Audio Anomaly Detection Using Feature Reconstruction

This study proposes an anomaly detection method based on feature reconstruction from subsampled audio signals. It shows that analyzing reconstructed features helps in identifying irregularities in audio, making it useful for detecting manipulated or synthetic speech.

7.Javed R. et al. (2021) – Digital Video Forensics Survey

Although focused on video, this research provides insights into deepfake detection challenges and methodologies. It explains that similar techniques can be adapted for audio deepfake detection using feature extraction and classification models.

8.**Stupp C. (2019) – AI Voice Cloning in Cybercrime Case**

This real-world case study reports how fraudsters used AI-based voice cloning to mimic a CEO's voice and commit financial fraud. It demonstrates the serious risks of deepfake audio and the need for reliable detection systems in security applications.

9.**Stupp C. (2019) – AI Voice Cloning in Cybercrime Case**

This real-world case study reports how fraudsters used AI-based voice cloning to mimic a CEO's voice and commit financial fraud. It demonstrates the serious risks of deepfake audio and the need for reliable detection systems in security applications.

10.RapidMiner (2018)–Feature Selection and Optimization Techniques

This work focuses on feature engineering techniques in machine learning. It highlights that proper feature extraction methods such as MFCC and CQCC significantly improve model accuracy, which is essential for detecting deepfake audio effectively.

III. Methodology

3.1 Existing System

Audio deepfake detection in existing systems primarily involves analyzing speech signals using traditional signal processing techniques and basic machine learning models. Most systems utilize spectral analysis methods such as frequency and waveform examination to identify irregularities in audio signals. Some approaches employ machine learning algorithms and basic neural networks using limited feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC).

These systems attempt to differentiate between real and synthetic audio by analyzing variations in pitch, tone, and frequency patterns. However, traditional methods face significant challenges in detecting advanced deepfake audio generated using modern artificial intelligence techniques, especially when the synthetic audio closely resembles natural human speech.

Low Accuracy in Deepfake Detection: Existing systems often fail to accurately detect highly sophisticated deepfake audio generated using advanced AI models. As deepfake technology evolves, the synthetic audio becomes more realistic, making it difficult for traditional detection methods to identify subtle differences between real and fake signals.

Lack of Real-Time Processing: Many existing approaches are not designed for real-time analysis and require significant processing time. This delay makes them unsuitable for applications where immediate detection is necessary, such as fraud prevention and live communication monitoring.

Single Algorithm Dependency: Most systems rely on a single detection algorithm, which limits their overall performance and reliability. The use of only one model reduces the system's ability to capture diverse audio characteristics, leading to lower accuracy and increased chances of misclassification.

High False Prediction Rates: Existing methods often produce higher rates of false positives and false negatives due to insufficient feature extraction and limited model capability. This reduces the trust and effectiveness of the detection system in practical scenarios.

Lack of User-Friendly Interface: Many existing systems do not provide an interactive or user-friendly interface, making them difficult to use for non- technical users. This limits their accessibility and real-world applicability.

3.2 Proposed System

The proposed system focuses on developing an efficient and accurate Audio Deepfake Detection System using advanced machine learning techniques. This system utilizes a dual-algorithm approach combining Support Vector Machine (SVM) with Mel-Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Model (GMM) with Constant-Q Cepstral Coefficients (CQCC) for robust audio analysis. The input audio is first preprocessed through normalization, resampling, and padding to ensure consistency. Feature extraction is then performed to capture important spectral and frequency characteristics of the audio signal.

These extracted features are passed to trained classification models, which analyze and classify the audio as real or fake. The system also incorporates a web-based interface using Streamlit, allowing users to upload audio files and receive real-time detection results along with confidence scores and visual representations such as waveform and spectrogram analysis. This approach improves detection accuracy, efficiency, and usability compared to existing systems.

Advantages of Proposed System

High Detection Accuracy: The use of a dual-algorithm approach combining
SVM and GMM significantly improves the accuracy of detecting real and fake audio compared to traditional single-model systems.

Real-Time Processing: The system is designed to process audio inputs quickly and provide instant results, making it suitable for real-time applications such as fraud detection and media verification.

Robust Feature Extraction: Advanced feature extraction techniques like
MFCC and CQCC effectively capture both spectral and frequency characteristics of audio signals, enhancing classification performance.

Reduced False Predictions: By using multiple algorithms and better feature representation, the system minimizes false positives and false negatives, increasing reliability.

User-Friendly Interface: The Streamlit-based web application provides an easy-to-use interface where users can upload audio files and view results without technical expertise.

Visualization Support: The system provides graphical representations such as waveform, spectrogram, and log-likelihood analysis, helping users better understand the detection process.

3.3 Feasibility Study

The feasibility study is an important step in system development that determines whether the proposed system is practical and beneficial. It evaluates the system based on technical, economic, and operational aspects to ensure successful implementation.

3.3.1 Economic Feasibility

The proposed system is highly economical as it is developed using open-source technologies and libraries. Tools such as Python, Streamlit, and Scikit-learn are freely available, eliminating the need for expensive software licenses. This significantly reduces the overall development cost.The hardware requirements are also cost- effective, as the system can run on standard computers without the need for high-end infrastructure. There are minimal maintenance and operational costs involved, as updates and improvements can be made using freely available resources.
Additionally, the system provides high value in terms of its applications in security, media verification, and fraud detection, making it a cost-efficient solution. Therefore, the project is financially viable and suitable for implementation within a limited budget.

3.3.2 Technical Feasibility

The proposed Audio Deepfake Detection System is technically feasible as it is built using well-established and widely supported technologies such as Python, Streamlit, Librosa, NumPy, and Scikit-learn. These tools provide efficient support for audio processing, feature extraction, and machine learning model development. The system utilizes algorithms like Support Vector Machine (SVM) and Gaussian Mixture Model (GMM), which are proven techniques for classification and statistical modeling tasks.

3.3.3 Social Feasibility

The proposed Audio Deepfake Detection System is socially beneficial as it addresses the growing concern of misuse of synthetic audio in society. With the increasing use of deepfake technology, there is a high risk of voice-based fraud, misinformation, and identity theft.

This system helps in identifying fake audio, thereby promoting trust and authenticity in digital communication.The system can be effectively used in areas such as media verification, law enforcement, and cybersecurity to prevent the spread of false information and protect individuals from fraudulent activities. It contributes to creating a safer digital environment by reducing the impact of malicious audio content.

Additionally, the system is designed to be user-friendly and accessible, allowing people from different backgrounds to use it without requiring technical expertise. This increases its acceptance and usability among the general public.

3.4 System Specifications

3.4.1Hardware Specifications

Processor: Intel Core i3 or higher

RAM: Minimum 4 GB (8 GB recommended for better performance)
Storage: At least 2 GB of free disk space
System Type: 64-bit system
Input Devices: Keyboard and mouse
Output Devices: Monitor or display screen

3.4.2 Software Specifications

Operating System: Windows, Linux, or macOS
Programming Language: Python 3.8 or higher
Development Environment: VS Code / Jupyter Notebook / PyCharm
Libraries and Frameworks:
Streamlit (for web interface)
Librosa (for audio processing)
Scikit-learn (for machine learning models)
NumPy (for numerical computations)
Matplotlib (for visualization)
Joblib (for model loading)
Web Browser: Google Chrome, Mozilla Firefox, or Microsoft Edge

3.5 General Architecture

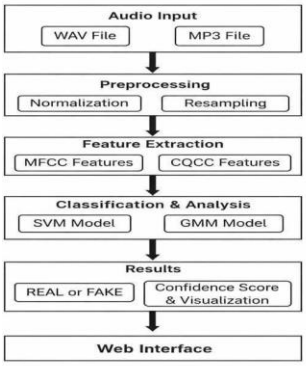

Fig 3.5: General Architecture

The above diagram represents the overall architecture of the Audio Deepfake Detection System. The process begins with audio input in the form of WAV or MP3 files, which are then passed through a preprocessing stage that includes normalization and resampling to ensure consistency in the data. Next, feature extraction is performed using MFCC and CQCC techniques to capture the essential spectral and frequency characteristics of the audio signal. These features are then analyzed using classification models, namely Support Vector Machine (SVM) and Gaussian Mixture Model (GMM), to determine whether the audio is real or fake.
The system then generates results along with confidence scores and visualizations to provide better insights into the analysis. Finally, all outputs are displayed through a user-friendly web interface, enabling real-time interaction and easy accessibility for users.

3.6 Design Phase

3.6.1 Data Flow Diagram

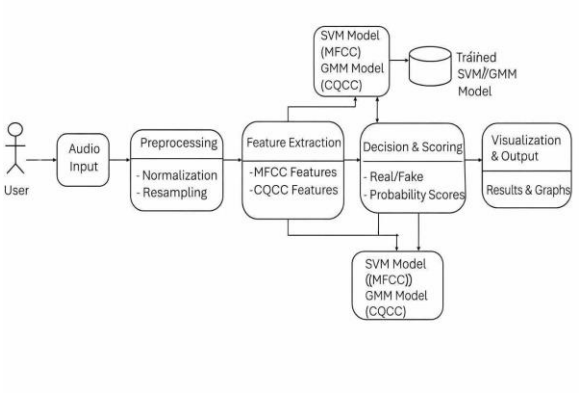

Fig 3.6.1: Data Flow Diagram

This data flow diagram illustrates a complete pipeline for audio-based authenticity detection, where the system determines whether a given audio sample is real or fake. The process starts with the user providing an audio input, which is then subjected to preprocessing techniques such as normalization and resampling to remove noise and standardize the signal format.
After preprocessing, the system performs feature extraction, where key acoustic features like MFCC (Mel-Frequency Cepstral Coefficients) and CQCC (Constant Q Cepstral Coefficients) are derived to capture the unique characteristics of the audio signal.
These extracted features are then fed into trained machine learning models, specifically Support Vector Machine (SVM) and Gaussian Mixture Model (GMM), which have been previously trained using labeled data.

3.6.2 Usecase Diagram

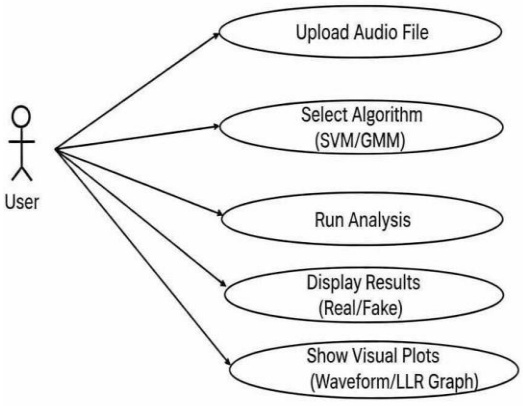

Fig3.6.2:UseCase Diagram

The above diagram represents the use case of the Audio Deepfake Detection System, showing the interaction between the user and the system. The user begins by uploading an audio file and selecting the desired algorithm, either SVM or GMM. The system then performs the analysis on the uploaded audio and processes it using the selected model. After processing, the system displays the result indicating whether the audio is real or fake. Additionally, it provides visual representations such as waveform and log-likelihood ratio graphs to help the userbetterunderstandthe analysis.

This interaction ensures a simple and user-friendly experience. Furthermore, the system is designed to ensure smooth interaction by guiding the user through each step of the process. It allows flexibility in choosing different algorithms for comparison, which enhances the reliability of the results. The visual outputs provided by the system help in better interpretation of the audio characteristics .

3.6.3 Class Diagram

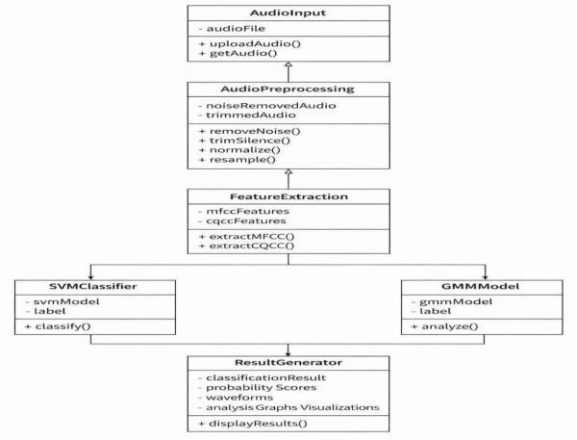

Fig 3.6.3: Class Diagram

The above diagram represents the class structure of the Audio Deepfake Detection System. The AudioInput class handles uploading and retrieving audio files, which are then processed in the AudioPreprocessing class through operations like noise removal, normalization, and resampling. The processed audio is passed to the FeatureExtraction class, where MFCC and CQCC features are extracted. These features are analyzed by two models: SVMClassifier for classification and GMMModel for statistical analysis. Finally, the ResultGenerator class combines the outputs and displays the classification result, probability scores, and visualizations to the user.

3.6.4 Activity Diagram

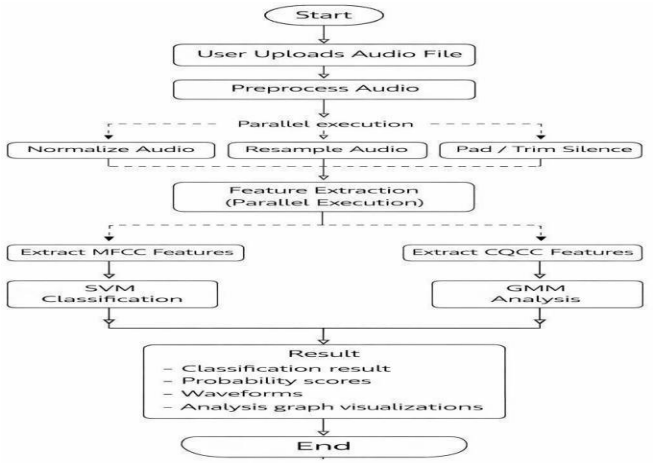

Fig 3.6.4: Activity Diagram

The above diagram illustrates the workflow of the Audio Deepfake Detection System from start to end. The process begins when the user uploads an audio file, which is then preprocessed to ensure consistency. During preprocessing, operations such as normalization, resampling, and padding or trimming of silence are performed in parallel to improve efficiency. After preprocessing, feature extraction is carried out in parallel using two techniques: MFCC and CQCC, which capture different characteristics of the audio signal. The extracted MFCC features are passed to the SVM model for classification, while the CQCC features are analyzed using the GMM model.

Both models work simultaneously to evaluate the authenticity of the audio. Finally, the system generates the output, including classification results (real or fake), probability scores, waveform representations, and analysis visualizations, before ending the process.

3.6.5 Sequence Diagram

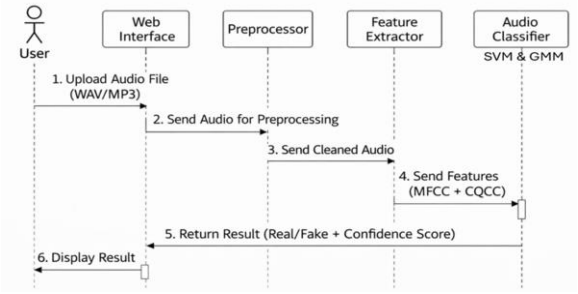

Fig 3.6.5: Sequence Diagram

The above diagram represents the overall architecture of the Audio Deepfake Detection System. The process begins with audio input in the form of WAV or MP3 files, which are then passed through a preprocessing stage that includes normalization and resampling to ensure consistency in the data. Next, feature extraction is performed using MFCC and CQCC techniques to capture the essential spectral and frequency characteristics of the audio signal.

These features are then analyzed using classification models, namely Support Vector Machine (SVM) and Gaussian Mixture Model (GMM), to determine whether the audio is real or fake. The system then generates results along with confidence scores and visualizations to provide better insights into the analysis. Finally, all outputs are displayed

through a user-friendly web interface, enabling real-time interaction and easy accessibility for users.

IV. Results

SVM(MFCC)

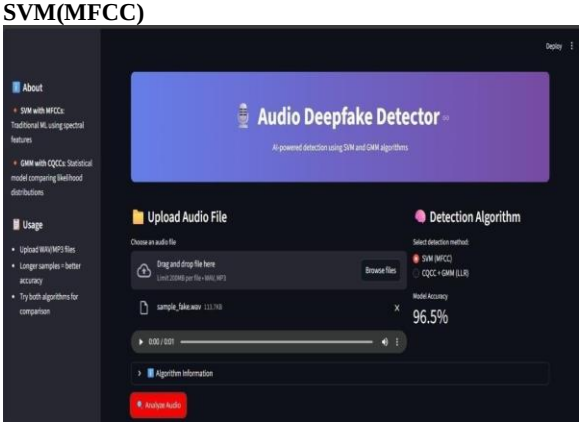

Fig 4.1 : Uploading Audio File

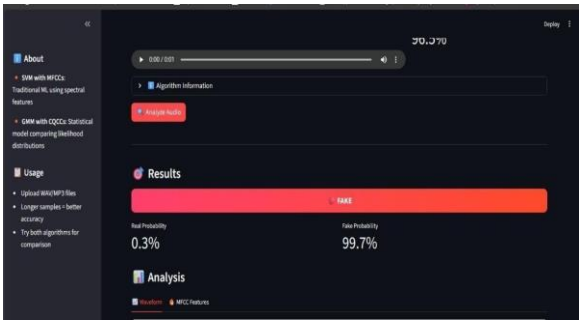

Fig 4.2 : Analyzing Audio File

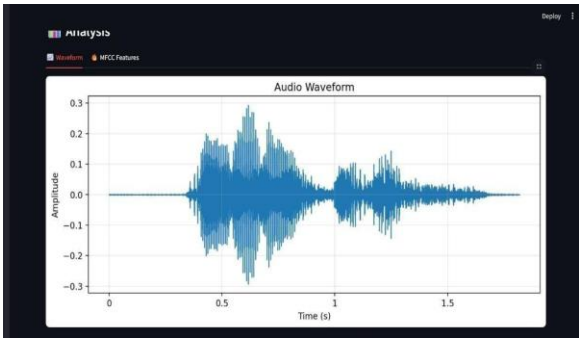

Fig 4.3 : Audio Waveform

Fig 4.4 : MFCC Features

Analysis of Output Results (SVM-MFCC):
The figures above present a comparative look at the system's behaviour when processing both Real (Human) and Fake (Synthetic) audio samples using the SVM-MFCC pipeline.

For Real Audio: The MFCC Feature Analysis displays a rich, non-linear spectral distribution. These patterns represent the natural resonance and complex fluctuations of the human vocal tract. The SVM classifier identifies these as authentic, resulting in a "REAL" classification with high confidence. This confirms the model's ability to recognize the "warmth" and natural variability of human speech.

For Fake Audio: In contrast, the synthetic sample reveals distinct spectral regularities or "robotic" artifacts in the MFCC heat map. These are the digital "fingerprints" left behind by AI voice-cloning vocoders. The SVM algorithm successfully detects these anomalies as falling outside the human distribution, correctly flagging the sample as "FAKE" with a high probability (often reaching 99.9%).

V. Conclusion

Conclusion of Test Cases:
The successful detection in both scenarios validates the robustness of the MFCC-based feature extraction. By comparing the two, it is evident that while the raw audio might sound similar to the human ear, the underlying spectral analysis provides a definitive boundary for the Support Vector Machine to achieve its 96.5% accuracy in distinguishing synthetic voice clones from genuine human recordings.

GMM +CQCC(LLR)

Fig 5.1: Uploading Audio

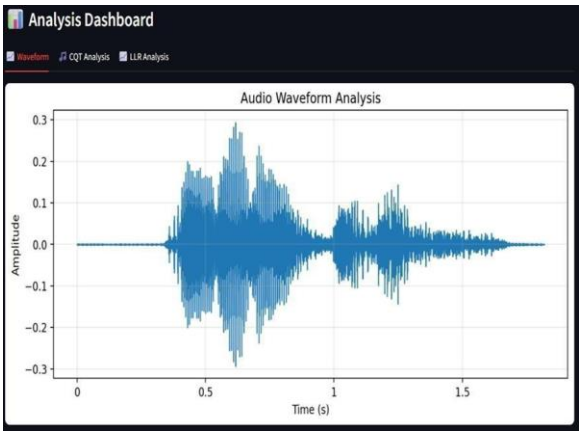

Fig 5.2 : Waveform

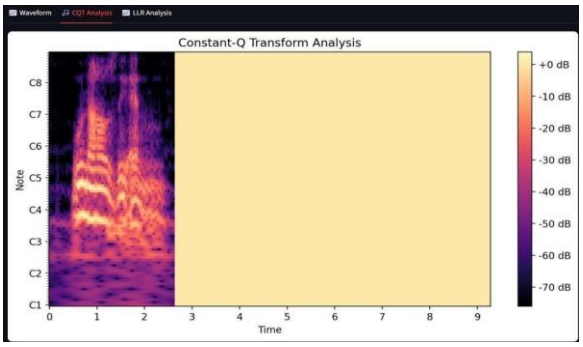

Fig 5.3: Constant-Q Transform Analysis

Fig 5.4: LLR Analysis

Analysis of GMM-CQCC Pipeline Results
The figures above present the end-to-end processing of an audio sample using the GMM (Gaussian Mixture Model) combined with CQCC (Constant Q Cepstral Coefficients). Unlike the discriminative nature of SVM, this pipeline uses a generative statistical approach to identify synthetic voices.

Feature Extraction (CQT): The process begins with a Constant-Q Transform, which provides a high-resolution spectral mapping across musical octaves. As seen in the CQT heat map, this allows the system to capture the fine-grained harmonic structures of the voice. Synthetic vocoders often leave subtle mathematical "footprints" in these harmonics that standard analysis might miss.

Statistical Decision (LLR): The extracted CQCC features are then processed through the Log-Likelihood Ratio (LLR) analysis. The LLR graph provides a frame-by- frame breakdown of the model's "internal thinking." By plotting values against the Decision Boundary (LLR=0), the system visualizes whether specific segments of the audio align more closely with "Real" (Human) or "Fake" (Synthetic) Gaussian clusters. Final Classification: By aggregating these frame-level LLR scores, the system calculates a final probability percentage. For a real sample, the model typically yields a dominant Real Probability (e.g., 58.0%), while for a synthetic sample, the Fake Probability will prevail.

Conclusion of the GMM Pipeline:
While the GMM-CQCC model operates with a 90.4% accuracy, its strength lies in its high- resolution spectral sensitivity. It acts as a vital secondary validator in this project, ensuring that even high-quality voice clones—which might bypass simpler filters—are caught by their statistical inconsistencies in the harmonic domain.

The implementation of the Deep Fake Audio Detection System using SVM with MFCC and GMM with CQCC demonstrates a significant advancement in identifying synthetic speech. The dual-algorithm approach allows the system to capture both broad spectral features and high-resolution frequency details, making it highly effective at distinguishing between real human voices and sophisticated AI-generated clones.

By leveraging machine learning and advanced signal processing, the system efficiently classifies audio files even in challenging conditions. The high accuracy levels—96.5% for the SVM model and 90.4% for the GMM model—confirm the system's effectiveness. Additionally, the real-time Streamlit web interface provides an intuitive and accessible platform for users to verify audio authenticity with visual support from waveforms and LLR analysis plots.

While the system performs exceptionally well on the current dataset, further scalability can be achieved by incorporating Deep Learning architectures like RNNs and expanding the training data to include a wider variety of languages and background noise environments. This project provides a practical and reliable tool for sectors like digital forensics, financial security, and media verification, ensuring trust in digital communications.

VI. Future Work

While the current system demonstrates high accuracy using SVM and GMM models, several advancements can further improve its robustness and scalability: Implementation of Deep Learning: Transitioning from traditional machine learning to deep learning architectures, such as Recurrent Neural Networks (RNN/LSTM) for temporal pattern recognition, could significantly improve detection of highly sophisticated "live" voice clones

Expansion of Datasets: Training the models on a more diverse and larger dataset (like ASVspoof 2021) that includes multiple languages, accents, and various emotional tones would enhance the system's global applicability.

Robustness to Noise: Future versions can integrate Adaptive Noise Cancellation or speech enhancement preprocessing to maintain high accuracy when analyzing audio recorded in noisy environments like public streets or over low-quality phone lines.

Cross-Domain Generalization: Enhancing the model to generalize better across different recording devices (microphones vs. smartphones) would make it more reliable for real-world forensic applications.

Real-Time Streaming Analysis: Developing the capability to analyze live streaming audio (e.g., during a Zoom or WhatsApp call) would provide immediate protection against real-time voice-cloning fraud.

VII. References

1.Abbasi, A. R., Javed, A., Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, ''A largescale benchmark dataset for anomaly detection and rare event classification for audio forensics,'' IEEE Access, vol. 10, pp. 38885–38894, 2022.

2.R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, ''A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions,'' IEEE Access, vol. 10, pp. 11065–11089, 2022.

3.R. Javed, Z. Jalil, W. Zehra, T. R. Gadekallu, D. Y. Suh, and M. J. Piran, ''A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions,'' Engineering Applications of Artificial Intelligence, vol. 106, Nov. 2021, Art. no. 104456.

4.Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, ''Privacy of web browsers: A challenge in digital forensics,'' in Proc. Int. Conf. Genetic Evol.
Comput., Springer, 2021, pp. 493–504.

5.S. Anwar, M. O. Beg, K. Saleem, Z. Ahmed, A. R. Javed, and U. Tariq, ''Social relationship analysis using state-of-the-art embeddings,'' ACM Trans. Asian LowResource Lang. Inf. Process., Jun. 2022.

6.J. Stupp, ''Fraudsters used AI to mimic CEO's voice in unusual cyber- crime case,'' Wall Street Journal, vol. 30, no. 8, pp. 1–2, 2019.

7.S. Ahmed, Z. A. Abbood, H. M. Farhan, B. T. Yasen, M. R. Ahmed, and A.
D. Duru, ''Speaker identification model based on deep neural networks,'' Iraqi Journal of Computer Science and Mathematics, vol. 3, no. 1, pp. 108–114, Jan.
2022.

9.J. K. Hansen, P. C. Hansen, and S. Holdt, ''Deep learning for audio forensics: A review of challenges and emerging solutions,'' IEEE Signal Processing Magazine, vol. 39, no. 5, pp. 85–97, Sept. 2022.

10.M. Todisco, H. Delgado, and N. Evans, ''Constant Q cepstral coefficients for spoofing detection,'' Computer Speech & Language, vol. 45, pp. 516–535, Sept.
2017.