International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Harmonizing Algorithms: An Approach to Enhancing Audio Deepfake Detection

Shwetambari Borade[1], Nilakshi Jain[2], Bhavesh Patel[3], Vineet Kumar[4], Yash Nagare[5], Shubham Kolaskar[6], Jayan Shah [7], Pratham Shah[8], Mustansir Godhrawala[9]

Submitted: 25/01/2024 **Revised:** 03/03/2024 **Accepted:** 11/03/2024

Abstract: This research aims to enhance the detection of audio deepfakes by developing a real-time, highly accurate methodology that addresses existing technological and ethical gaps in the field. Employing advanced algorithms for feature extraction, the study innovatively utilizes a multifaceted approach by integrating an MFCC-based SVM classifier, which achieved a remarkable 97.28% accuracy, and a Neural Network with attention mechanisms, with a 91.04% accuracy rate. A novel aspect of our methodology is the use of multiple models in tandem to verify the authenticity of input audio, significantly boosting the reliability of detection. Leveraging the 'For-Original' dataset for exhaustive training and validation, our methods have shown exceptional effectiveness in distinguishing genuine audio from synthetic counterparts. These findings not only demonstrate significant improvements in existing deepfake detection techniques but also introduce a novel approach to comparative model analysis. This contribution is pivotal in advancing the field of digital media integrity, offering new avenues for ensuring the authenticity of audio content in the era of sophisticated digital forgeries.

Keywords: *Audio Deepfake Detection, Comparative Model Verification, Ethical Audio Forensics, Real-time Speech Authenticity, SVM-Neuron Network Fusion*

1 Assistant Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0000-0001-7547-6351
2 Professor Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0000-0002-6480-2796
3 Professor, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0001-0363-9809
4 Founder & Global President, CyberPeace Foundation, Delhi, India
ORCID ID : 0009-0000-3806-7380
5 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0003-1266-3709
6 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0002-1394-7992
7 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0000-9677-9175
8Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0006-0935-6865
9 Student, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
ORCID ID : 0009-0005-4065-4361
** Corresponding Author Email: shwetambari.borade@sakec.ac.in*

1. Introduction

The rapid advancement of artificial intelligence and machine learning technologies has ushered in a new era of digital creativity and innovation, particularly in the generation of realistic audio and video content. Among these developments, audio deepfakes—synthetic audio recordings crafted to mimic real human voices with high accuracy—have emerged as a significant area of interest and concern. Recent developments in this field have demonstrated the capability to produce highly convincing fake audio content, raising critical issues related to security, privacy, and the dissemination of misinformation. Despite the remarkable progress in creating realistic synthetic audio, the technology's potential for misuse necessitates robust detection mechanisms to safeguard against deceptive practices and uphold digital media integrity.

The existing literature on deepfake detection primarily focuses on visual deepfakes, with less emphasis on the auditory aspects, indicating a gap in research dedicated to identifying and mitigating the risks posed by audio forgeries. This gap highlights the urgent need for dedicated research efforts towards developing effective and reliable audio deepfake detection systems. The present work is undertaken to address this need, offering a comprehensive analysis of current challenges, and proposing innovative solutions to enhance the detection of audio deepfakes. By employing a multifaceted approach that integrates an MFCC-based SVM classifier and a Neural Network with

attention mechanisms, our study aims to significantly improve the accuracy and reliability of existing detection methods.

Our research is among the first to utilize a comparative model verification approach, where multiple models are employed in tandem to validate the authenticity of input audio. This novel methodology not only contributes to the field by enhancing detection capabilities but also paves the way for future studies to explore and refine this approach further. The objectives of our study are twofold: to develop a highly accurate methodology for real-time audio deepfake detection and to bridge the current technological and ethical gaps identified in the literature.

By detailing the recent advancements in audio deepfake generation and detection, our work aims to inspire further investigation into this pressing issue, offering valuable insights and tools for researchers, technologists, and policymakers to combat the challenges posed by audio deepfakes in the digital age.

2. Literature Review

This study [1] underscores AI's dual role in creating and detecting deepfakes, highlighting progress in generating sophisticated deepfakes and the advancements in their detection. It identifies the need for real-time deepfake detection methods and the ethical considerations surrounding AI's application in digital deception. The research [2] explores various techniques for generating deepfakes in audio, including the use of recurrent neural networks. The study calls for the development of robust detection methods and comprehensive datasets to train and evaluate detection models effectively. The study [3] focuses on using MFCCs for feature extraction in fake voice detection, utilizing machine and deep learning for classification. There's a need for enhanced detection methods that can keep pace with the sophistication of synthetic audio technologies.

AVoiD-DF [4] integrates audio-visual cues to detect multi-modal forgeries, demonstrating effectiveness across various datasets. The study acknowledges the need for scalable and robust detection methods to combat evolving deepfake manipulation techniques. BTS-E [5] uses breathing sound patterns to improve deepfake detection, showing significant improvement in classifier performance. The paper suggests exploring applicability across different datasets and robustness against advancing deepfake generation methods. This work [6] introduces a comprehensive system for detecting deepfake audio in group conversations, utilizing a combination of deep learning architectures. It highlights the necessity of evaluating performance across diverse datasets and enhancing adaptability to new deepfake techniques.

A novel bi-level optimization technique [7] enhances the accuracy of deepfake audio detection. The method's scalability and adaptability to different datasets and emerging deepfake threats require further exploration. The proposed SE-Res2Net-Conformer architecture [8] aims to improve the detection of synthetic voice and audio tampering. Future research is needed to assess performance across various scenarios and potential enhancements to improve detection accuracy. The study [9] evaluates RNNs and CNNs for identifying fake audio messages, proving highly efficient in a specific crisis dataset. There is a call for further assessment of the scalability and generalization of this approach across diverse situations and datasets. The survey [10] provides a systematic overview of audio deepfake detection, discussing techniques and challenges in the field. It indicates the necessity for future work to improve methods' generalization to unknown attacks and enhance interpretability of detection results.

3. Methodology

In Figure 3.1, we outline our methodical approach to audio signal analysis. The procedure commences with the collection of audio samples provided by users. These are processed through an 'Extract Audio Fingerprint' stage utilizing advanced algorithms to identify distinctive acoustic features. Post-extraction, the fingerprints are examined for correspondence within our dataset. Matches are channeled to a detailed analytical phase to assess match confidence and to distill further acoustic characteristics. The process culminates in the 'Visualization' stage, showcasing the data through waveforms, spectral views, and 3D spectrograms for in-depth interpretability. A more detailed explanation of each stage follows below, elucidating the intricacies of our methodology and the analytical techniques employed.

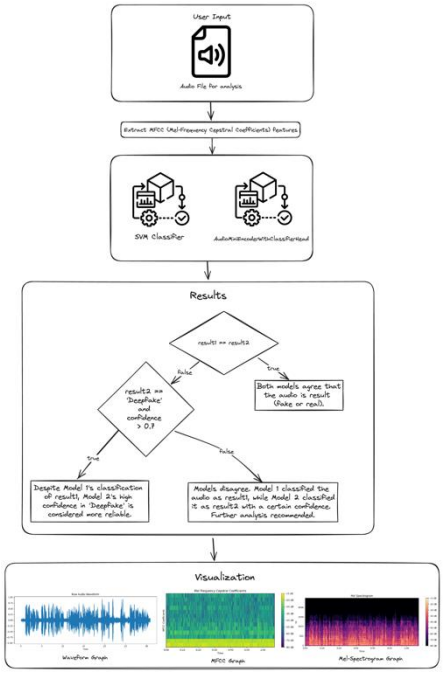

Fig 3.1: Proposed Architecture of our Audio Deepfake Detection System

3.1. Dataset Selection

For our research, we selected the 'for-original' variant of the Fake-or-Real (FoR) Dataset, sourced from the APTLY lab and accessible through the Biometric Intelligence Lab at York University. This dataset variant was chosen due to its comprehensive nature and meticulous curation, making it a suitable choice for training, and evaluating deepfake audio detection models.

1. **Variant Selection**: The 'for-original' variant was preferred for its pristine and unaltered nature, aligning with our system's design philosophy to train the model under conditions that closely resemble real-world scenarios.
2. **Volume and Diversity**: With over 195,000 audio utterances, the dataset offers a rich collection of speech variations, encompassing diverse vocal characteristics influenced by factors such as speaker identity, accent, and linguistic content.
3. **Source Inclusivity**: The dataset includes samples from both authentic human speech and synthetic speech outputs generated by advanced Text-to-Speech (TTS) technologies such as Deep Voice 3 and Google Wavenet TTS. Additionally, human speech samples are sourced from reputable datasets like Arctic, LJSpeech, and VoxForge, ensuring a comprehensive representation of speech types.
4. **Quality Assurance**: High fidelity recordings guarantee that the dataset maintains the integrity of acoustic properties present in both genuine and synthetic speech, enabling robust training and evaluation of the deepfake audio detection model.

By leveraging the 'for-original' variant of the FoR Dataset, our research benefits from a diverse and high-quality dataset that facilitates the development of a robust deepfake audio detection system capable of discerning between authentic and synthetic speech inputs effectively.

3.2. Feature Extraction

Model 1 employs Mel-Frequency Cepstral Coefficients (MFCCs) for the extraction of audio signal's spectral features. Recognized for their precision in isolating the distinct timbral characteristics of speech, MFCCs serve as a critical tool in our analysis, enabling the effective differentiation between authentic and fabricated audio clips. This method capitalizes on the inherent spectral properties embedded within the audio, leveraging the power of MFCCs to capture the essence of natural speech patterns, and identifying anomalies that suggest manipulation.

Model 2 advances our feature extraction capabilities by processing audio data through a series of convolutional layers, augmented with sophisticated attention mechanisms. This design is tailored to enhance the model's focus on audio segments most likely to contain traces of manipulation, thereby providing a more refined analysis. By considering the contextual and temporal dynamics present within the audio signal, this model uncovers subtle indicators of deepfake content. The integration of attention mechanisms not only improves the accuracy of detection but also highlights the model's ability to discern complex patterns of audio manipulation, making it a potent tool for identifying synthetic audio with high precision.

3.3. Model Training

3.3.1. Model 1: MFCC-based SVM Classifier

Our deepfake audio detection system leverages Mel-Frequency Cepstral Coefficients (MFCCs) in conjunction with a Support Vector Machine (SVM) classifier to discern between genuine and manipulated audio inputs. MFCCs are renowned for their ability to encode the timbral and textural aspects of sound, making them particularly suitable for speech and audio analysis tasks. The process of computing MFCCs involves several computational stages designed to transform raw audio waveforms into a feature set that faithfully captures essential spectral properties while aligning with the human auditory system's perceptive capabilities.

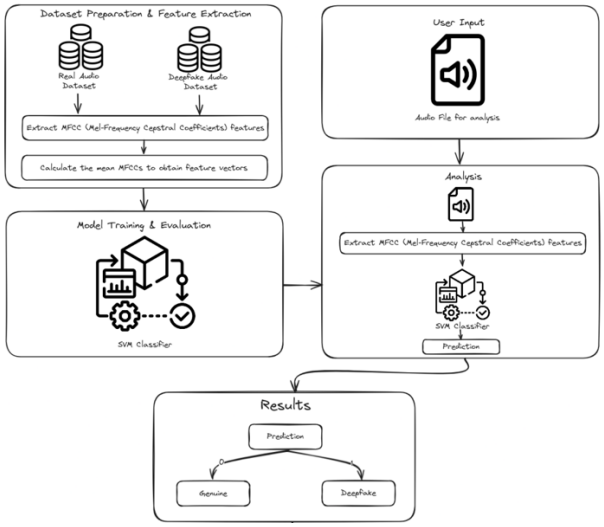

Fig 3.1: Proposed system architecture for model 1

3.3.1.1. Feature Extraction Process

1. Discrete Fourier Transform (DFT): The process begins by applying the DFT to the raw audio waveform, transforming it from the time domain into the frequency domain.
2. Mel Filter Bank: The power spectrum obtained from the DFT is then passed through a set of bandpass filters known as the Mel filter bank. These filters are spaced uniformly on the Mel scale, which mimics the nonlinear human perception of sound.

3. Logarithmic Scale: The log filter bank energies are calculated using a logarithmic scale to mimic the way human ears perceive loudness, producing precise measurements.

4. Discrete Cosine Transform (DCT): Finally, the log Mel filter bank energies undergo the DCT to calculate the MFCCs. This step decorrelates the log Mel spectrum and yields a compressed representation of the filter banks, emphasizing lower-order coefficients that capture the most salient aspects of the signal.

3.3.1.2. Model Training and Evaluation

Our model undergoes a rigorous training and evaluation process to guarantee optimal performance in distinguishing genuine from AI-generated audio. First, we meticulously standardize the feature set, ensuring all features contribute equally during training. This levels the playing field for each feature and prevents any from dominating the learning process. Next, the data is strategically split into training and testing sets using a stratified approach. This maintains a balanced class distribution across both sets, leading to a more robust evaluation.

The core of the model is a Support Vector Machine (SVM) classifier. During training, the SVM identifies hyperplanes that effectively create boundaries within the scaled training data, separating the genuine and deepfake audio data points. To assess the model's effectiveness, we leverage two key metrics. Accuracy provides a high-level overview of the model's success rate in correctly classifying audio samples.

For a deeper dive, we utilize a confusion matrix (refer to Fig. 3.3). This visual tool breaks down the model's performance across different categories. It details the number of true positives (correctly classified genuine audio), false negatives (genuine audio misclassified as deepfakes), false positives (deepfakes misclassified as genuine audio), and true negatives (correctly classified deepfakes). By analysing this breakdown, we can pinpoint areas for improvement and refine the model to minimize misclassifications. Finally, for efficient deployment and future analysis, the trained model and scaler are serialized using Joblib.

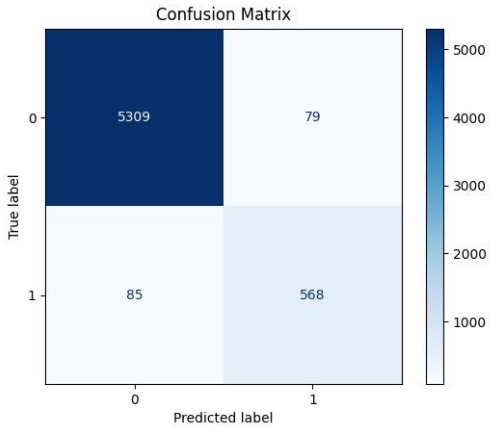

Fig 3.2: Confusion Matrix for model 1

3.3.2. Model 2: Neural Network with Attention Mechanism

The '**Audio Mini Encoder with Classifier Head**', represents a leap towards leveraging neural network architectures for deepfake detection. This model is distinguished by its use of convolutional layers, residual blocks, and, most notably, attention mechanisms that prioritize the analysis of specific segments of the audio signal deemed most relevant for classification purposes. The architecture begins with an initial convolutional layer that processes the raw audio input, followed by a series of residual blocks that enhance feature extraction through deep layers while preventing the vanishing gradient problem. Attention mechanisms further refine the model's focus, allowing it to discern subtle cues indicative of audio manipulation. This model's training involves the 'For-Original' dataset, enabling it to distinguish between genuine and AI-generated audio with a noted accuracy of 91.04%.

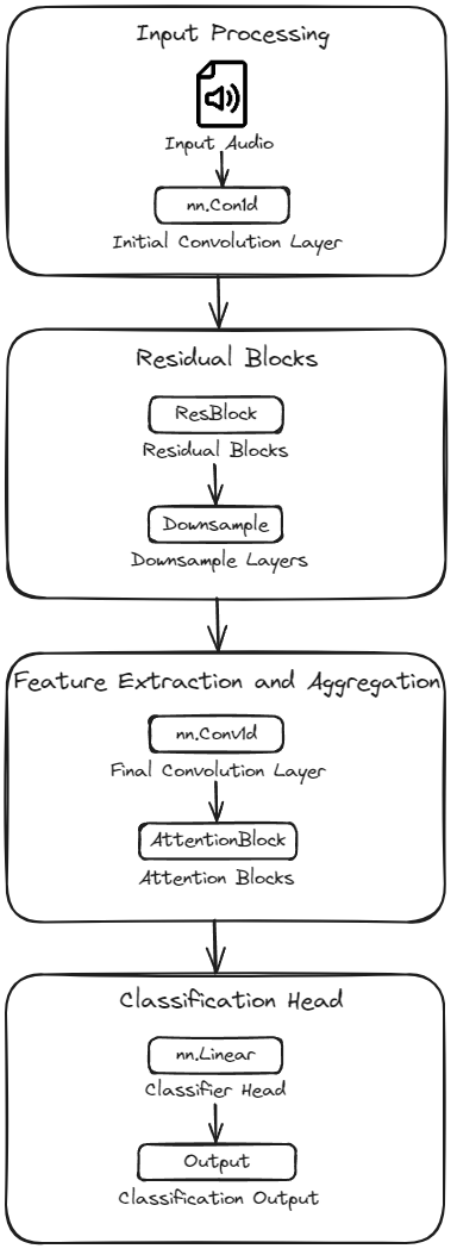

Fig 3.3: Proposed architecture of model 2

3.3.2.1. Architecture Overview

1. Input Processing: The audio input undergoes initial processing through a convolutional layer to transform its spectral dimensions.
2. Residual Blocks: Multiple residual blocks (ResBlocks) are employed to process features extracted by the initial convolution, facilitating the capture of complex patterns within the audio data.
3. Downsample Layers: Periodic downsampling between ResBlocks reduces spatial dimensions of feature maps, concentrating feature information and reducing computational load.
4. Feature Extraction and Aggregation: A final convolutional layer refines feature representation, followed by attention blocks to focus on relevant audio segments. The attention mechanism enhances sensitivity to subtle cues indicative of deepfake audio.
5. Classification Head: A linear layer acts as a classifier, mapping aggregated feature representation to output space and producing classification probabilities.

3.3.2.2. Model Evaluation

The pre-trained model's performance was evaluated using the 'For-Original' dataset, known for its realistic and diverse speech samples. This evaluation included a confusion matrix analysis (refer to Figure 3.5) and yielded an accuracy of 91.04%. The confusion matrix itself breaks down the model's classification results into four categories: true positives (genuine audio correctly classified), false negatives (genuine audio misclassified as deepfakes), false positives (deepfakes misclassified as genuine audio), and true negatives (deepfakes correctly classified). Analyzing these values provides a deeper understanding of the model's strengths and weaknesses in differentiating real from AI-generated speech. While high accuracy is promising, the confusion matrix can reveal areas for improvement, such as reducing false positives or false negatives. Continuous refinement through techniques like data augmentation and exploring new model architectures will be crucial to enhance the model's robustness and real-world applicability.

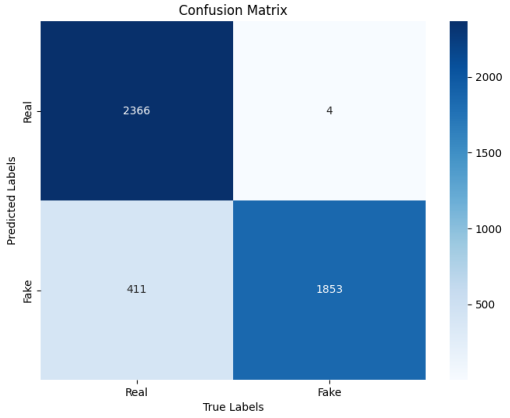

Fig 3.4: Confusion Matrix for Model 2

3.3.3. Model Comparison

The comparative analysis is a cornerstone of our methodology. Outputs from both models are evaluated using the *'compare_results'* function, which assesses the models' agreement and weighs the confidence score provided by Model 2. This comparison is crucial for resolving discrepancies and ensuring a more accurate final determination of the audio's authenticity.

3.3.3.1. Comparative Analysis Criteria

The '*compare_results'* function underpins our comparative framework, embodying a logic that refines the detection process:

1. Consensus-Based Verification: If both models classify

the audio similarly, their agreement is deemed reliable, and the consensus classification is accepted as the final verdict.

2. Confidence-Weighted Decision: In instances where Model 2 identifies the audio as 'Deepfake' with a confidence level exceeding 70%, its classification takes precedence. This prioritization is based on the advanced analytical capabilities of neural networks, especially when the model exhibits high certainty in its detection.

Further Analysis Recommendation: Disagreements, particularly when Model 2's confidence is below 70%, signal the need for additional scrutiny. This scenario acknowledges the inherent challenges in deepfake detection and suggests a nuanced approach, advocating for further analysis to reach a conclusive determination.

4. Results and Discussion

4.1. Results Analysis

Fig 4.1: Results for Test Sample Audio (Genuine)

In Figure 4.1, the classification results of the audio sample 'Krishna.wav' are presented. Model 1's MFCC-based analysis classified the sample as genuine, demonstrating the model's effectiveness in capturing and evaluating the spectral properties of the audio. Correspondingly, Model 2, which incorporates convolutional neural network layers and attention mechanisms, aligned with Model 1's assessment, also identifying the sample as genuine. The agreement between these diverse models underscores the robustness of our detection approach and adds a layer of validation to the authenticity of the audio sample in question.

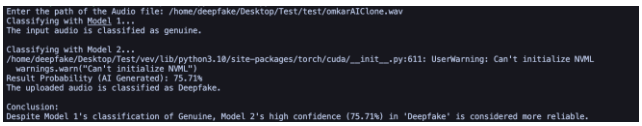

Fig 4.2: Results for Test Sample Audio (Fake)

In Figure 4.2, the analysis outcome for the 'omkarATClone.wav' audio file is documented. Model 1, employing MFCC feature extraction, classified the audio as genuine. However, Model 2, which is augmented with a convolutional neural network and attention mechanisms, assigned a high confidence level of 75.71% to the audio being a deepfake. This divergence in classification results highlights the intricate challenges involved in deepfake audio detection and demonstrates the necessity of leveraging multiple models to enhance the robustness and reliability of our detection methodologies.

4.2. Visualization Analysis

4.2.1. Waveform Analysis

Our analysis commences with a comparative visualization of audio waveforms, crucial for distinguishing between authentic and fabricated acoustic signals. These waveform plots serve as a fundamental tool to discern the nuanced discrepancies between legitimate and deepfake audio samples.

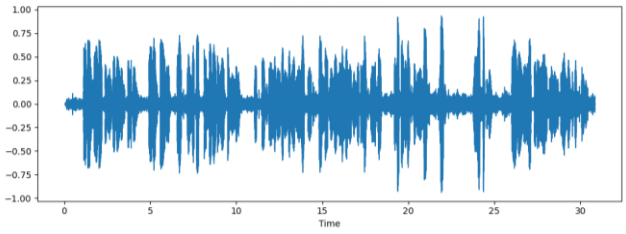

Fig 4.3: Waveform of Genuine Audio Sample

Figure 4.3 illustrates the waveform of an authentic audio track, labelled real_audio.wav, revealing the natural fluctuations and breadth of dynamic range typical of unaltered vocal recordings.

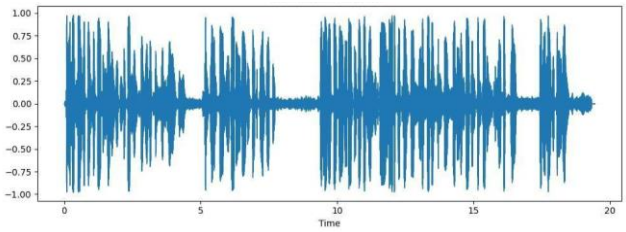

Fig 4.4: Waveform of Deepfake Audio Sample

In contrast, Figure 4.4 displays the waveform of a synthesized deepfake audio sample, serving as a key tool in detecting the synthesized patterns and anomalies characteristic of artificially generated speech.

4.2.2. Spectrogram Analysis

The spectrogram serves as a pivotal visualization tool within our detection methodology, offering a visual account of the audio signal's frequency spectrum over time. By implementing spectrogram analysis, we gain invaluable insight into the intricate frequency interactions within both original and manipulated audio samples. These spectrograms are integral to pinpointing the subtle yet distinct spectral discrepancies indicative of deepfake generation.

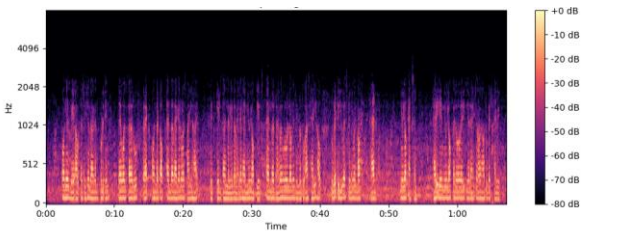

Fig 4.5: MelSpectrogram of Genuine Audio Sample

Figure 4.5 captures the Mel spectrogram of genuine audio,

illustrating the variable frequency modulations that signify the authenticity of human speech.

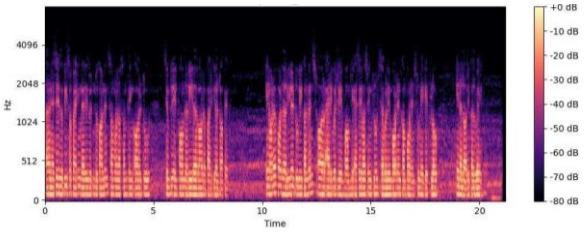

Fig 4.6: Melspectrogram of Deepfake Audio Sample

Figure 4.6, on the other hand, displays a Mel spectrogram for a piece of deepfake audio, characterized by atypical spectral patterns. The disparities observed, especially in the distribution of spectral energy, are critical for differentiating between synthesized and genuine speech.

4.2.3. MFCC Analysis

The Mel-Frequency Cepstral Coefficients (MFCCs) are of paramount significance in audio signal processing, especially for speech and audio recognition tasks. The MFCC plots encapsulate the audio signal's power spectrum, enabling the extraction of vital timbral characteristics that differentiate sounds and vocal tones. In detecting deepfakes, the MFCCs are instrumental in detecting the subtle deviations in speech patterns indicative of audio manipulation.

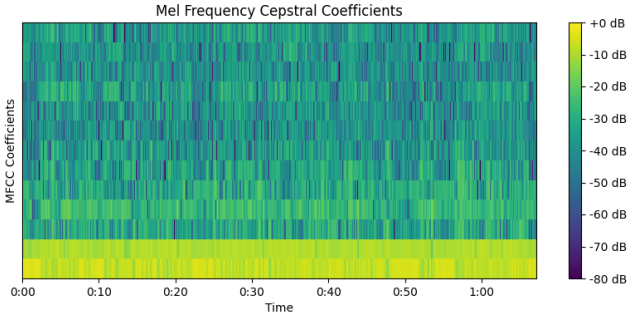

Fig 4.7: MFCC of Genuine Audio Sample

Figure 4.7 exhibits the MFCC visualization for a natural speech sample, where the cepstral features display a consistent pattern typical of genuine speech articulation.

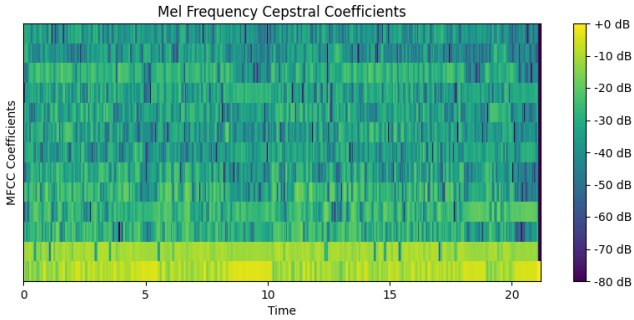

Fig 4.8: MFCC of Deepfake Audio Sample

Figure 4.8 delineates the MFCC visualization of synthetic speech, highlighting irregularities and deviations from the expected cepstral pattern, indicative of audio manipulation.

Through the side-by-side comparison of these visualization tools waveforms, spectrograms, and MFCC plots our analysis effectively identifies and elucidates the defining features of deepfake audio, bolstering the detection process.

5. Conclusion

In conclusion, our study aimed to address the pressing need for reliable audio deepfake detection in an era where synthetic media is rapidly advancing. By integrating and comparing the outcomes of an MFCC-based SVM classifier and a Neural Network with attention mechanisms, we not only achieved high accuracy rates (97.28% and 91.04%, respectively) but also presented a novel comparative analysis method. This dual-model approach enriches the current landscape of audio forensics, offering a more nuanced and reliable detection process for identifying synthetic speech.

Our research contributes fresh perspectives by demonstrating that while individual models are effective, a multi-faceted approach can significantly enhance the detection process. This is particularly evident when Model 2's attention mechanisms identify potential deepfakes with a high degree of confidence, even when Model 1 suggests authenticity.

However, study is not without its limitations. The difference in classification outcomes between the two models underscores the complex nature of audio deepfake detection and suggests areas for future research. Enhancing the sophistication of feature extraction methods and expanding the diversity of datasets for model training could address some of these challenges. Additionally, the exploration of real-time detection mechanisms would be an invaluable contribution to this field.

Future studies should aim to refine these models further, focusing on the adaptability of these methods to new and unforeseen types of audio manipulation. The development of a unified framework that encapsulates the strengths of various detection techniques could provide a more definitive solution to the challenges posed by audio deepfakes. The takeaway from our work is a compelling demonstration of the effectiveness of combining different analytical models, providing a robust tool for the ongoing battle against digital deception.

References

[1] M. A. Khder, S. Shorman, D. T. Aldoseri and M. M. Saeed, "Artificial Intelligence into Multimedia Deepfakes Creation and Detection," 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 2023, pp. 1-5, doi: 10.1109/ITIKD56332.2023.10099744.

[2] O. A. Shaaban, R. Yildirim and A. A. Alguttar, "Audio Deepfake Approaches," in IEEE Access, vol. 11, pp. 132652-132682, 2023, doi: 10.1109/ACCESS.2023.3333866.

[3] JH. H. Kilinc and F. Kaledibi, "Audio Deepfake Detection by using Machine and Deep Learning," 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), Istanbul, Turkiye, 2023, pp. 1-5, doi: 10.1109/WINCOM59760.2023.10323004.

[4] W. Yang et al., "AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2015-2029, 2023, doi: 10.1109/TIFS.2023.3262148.

[5] T. -P. Doan, L. Nguyen-Vu, S. Jung and K. Hong, "BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095927.

[6] R. L. M. A. P. C. Wijethunga, D. M. K. Matheesha, A. A. Noman, K. H. V. T. A. De Silva, M. Tissera and L. Rupasinghe, "Deepfake Audio Detection: A Deep Learning Based Solution for Group Conversations," 2020 2nd International Conference on Advancements in Computing (ICAC), Malabe, Sri Lanka, 2020, pp. 192-197, doi: 10.1109/ICAC51239.2020.9357161.

[7] M. Li, Y. Ahmadiadli and X. -P. Zhang, "Robust Deepfake Audio Detection via Bi-Level Optimization," 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), Poitiers, France, 2023, pp. 1-6, doi: 10.1109/MMSP59012.2023.10337724.

[8] L. Wang, B. Yeoh and J. W. Ng, "Synthetic Voice Detection and Audio Splicing Detection using SE-Res2Net-Conformer Architecture," 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, Singapore, 2022, pp. 115-119, doi: 10.1109/ISCSLP57327.2022.10037999.

[9] A. Khovrat and V. Kobziev, "Using Recurrent and Convulation Neural Networks to Indentify the Fake Audio Messages," 2023 IEEE 7th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), Kyiv, Ukraine, 2023, pp. 174-177, doi: 10.1109/MSNMC61017.2023.10329236.

[10] Yi, Jiangyan & Wang, Chenglong & Tao, Jianhua & Zhang, Xiaohui & Zhang, Chu & Zhao, Yan. (2023). Audio Deepfake Detection: A Survey.