Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 45 No. 1
January 2024

Enhancing Audio Deepfake Detection using Support Vector Machines and Mel-Frequency Cepstral Coefficients

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

[1] Professor, Cyber Security, Shah & Anchor Kutchhi Engineering College, Mumbai, India
[2] Assistant Professor, Cyber Security, Shah & Anchor Kutchhi Engineering College, Mumbai, India
[3] Professor, Computer Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India
[4] Founder & President, Cyber Peace Foundation, Delhi, India
[5,6,7,8,9] Student, Cyber Security, Shah & Anchor Kutchhi Engineering College, Mumbai, India

Abstract
This paper presents a machine learning system designed to differentiate real from synthetic speech using a Support Vector Machine (SVM) classifier. Trained on the 'for-original' Fake-or-Real (FoR) dataset, which consists of over 195,000 genuine and computer-generated utterances, the system uses Mel Frequency Cepstral Coefficients (MFCCs) to extract features. Evaluation results show a promising accuracy of 97.28%, indicating the system's potential efficacy in real-world applications. The work lays the foundation for future improvements in detection robustness and reliability by highlighting the significance of raw data in classifier training for deepfake detection.

Keywords: Deepfake Detection, Mel-Frequency Cepstral Coefficients (MFCCs), Support Vector Machine (SVM), Ethical Considerations, Audio Analysis, Real-world Applicability, Scalability Challenges, Responsible Technology Deployment, Media Manipulation, Feature Extraction.

1. Introduction

Deepfakes, manipulated media that realistically replace or alter a person's speech or appearance, have grown more and more problematic in the current digital era. Their ability to deceive audiences and spread misinformation poses significant threats to individual privacy, social trust, and even national security. While visual deepfakes have received much attention, audio-based deepfakes, often overlooked, can be equally impactful, manipulating speech content and impersonating voices with alarming accuracy. This makes reliable audio deepfake detection a critical challenge.

Existing research on deepfake detection has primarily focused on visual analysis, leveraging techniques like facial recognition and anomaly detection. However, these methods are often vulnerable to manipulation and may struggle with subtle audio changes. Audio-based detection, on the other hand, offers a promising alternative by analyzing the intrinsic characteristics of speech signals. This approach can potentially detect deepfakes based on subtle alterations in voice timbre, pitch, and pronunciation, even when the visual content appears unaltered.

The main focus of this research is to investigate how effectively we can identify audio based deepfakes by using Mel Frequency Cepstral Coefficients (MFCCs) and Support Vector Machine (SVM). MFCCs are a powerful feature extraction technique commonly used in audio

analysis, capturing the spectral characteristics of sound and providing a robust representation of speech signals. SVMs are highly regarded for their aptitude in tackling intricate data and attaining remarkable classification precision. Utilizing a combination of these methods, our objective is to create an effective and streamlined deepfake detection model that employs audio cues to distinguish authentic speech from manipulated recordings with exceptional accuracy.

This research will contribute to the growing field of deepfake detection by:

- Exploring the potential of audio-based analysis for deepfake identification.
- Developing and evaluating a robust audio deepfake detection model using MFCCs and SVMs.
- Exploring the advantages and drawbacks of using audio-based techniques in comparison, to the methods.
- Contributing to the development of effective tools and techniques for mitigating the harms of deepfakes.

The successful implementation of this research could lead to the development of reliable audio deepfake detection tools that can be integrated into various applications, such as social media platforms, news outlets, and even forensic investigations. This, in turn, can help combat the spread of misinformation, protect individual privacy, and promote trust in digital communication.

2. Literature Survey

The Authors in the paper titled [1] offer a comprehensive overview of the creation and detection of deepfakes, particularly focusing on multimedia content that encompasses audio and video elements. The authors delve into various deepfake generation techniques, such as voice cloning, lip-syncing, and facial manipulation, highlighting the growing sophistication of artificial intelligence in this domain. While the paper discusses the use of AI algorithms to analyze audio and visual features for deepfake detection, it does not provide detailed insights into specific detection models or their performance. Crucially, this study underscores the pressing need for robust detection methods and brings to light the broader challenges presented by deepfake technology. However, it falls short of providing an in-depth analysis or concrete results concerning audio-based detection methods, particularly those involving MFCCs and SVMs, which are central to our research.

This paper [2] delves into the diverse approaches used to produce these audio fakes. Through a detailed analysis, this paper provides a nuanced comprehension of the intricate methods employed in manipulating audio. The authors categorize and dissect audio deepfake methods based on their foundational principles, covering voice conversion (both text-to-speech and speech-to-speech), speech synthesis using deep learning models, voice cloning for high-fidelity synthetic speech production, and audio editing and splicing techniques that involve altering pitch and tempo or combining different audio segments. The paper's findings are significant, revealing that voice conversion and speech synthesis have seen considerable advancements, leading to more realistic fake audio. However, voice cloning is identified as a particularly challenging area, often necessitating extensive training data from the target speaker. The study also notes that while audio editing and splicing are effective in creating deepfakes, these techniques can lead to inconsistencies and increased detectability due to unnatural transitions. These insights provide a comprehensive backdrop for understanding the complexity of audio deepfake creation, highlighting the

technological progress and the persisting challenges in this field.

This paper, titled [3] offers valuable insights into the efficiency of machine and deep learning models, in identifying deepfakes. They conduct an analysis of machine learning algorithms, such as Support Vector Machines (SVMs) which have been successful in traditional classification tasks along with more advanced models, like Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs), which offer greater flexibility and complexity. These models are evaluated using different feature sets, such as raw audio waveforms and Mel-Frequency Cepstral Coefficients (MFCCs), the latter being particularly relevant to our study. The findings of this research are enlightening: ANNs and CNNs generally surpass SVMs in detecting audio deepfakes, exhibiting higher accuracy and sensitivity. Furthermore, the study highlights that MFCCs are superior features for this purpose, as they enable the models to discern subtle discrepancies more easily between authentic and manipulated audio. Notably, the paper underscores the efficacy of hybrid approaches that combine various models and features, suggesting that leveraging the strengths of multiple methods could lead to optimal performance in deepfake detection.

In the paper [4] the authors introduce 'POI-Forensics,' a novel approach to deepfake detection that significantly broadens the scope of traditional methods by focusing on combined audio-visual analysis to identify deepfakes targeting specific individuals or 'Persons-of-Interest' (POI). This innovative methodology employs contrastive learning to effectively distinguish between genuine and manipulated representations of a POI, contrasting similar and dissimilar examples to enhance the model's discrimination capabilities. Unique in its approach, POI-Forensics processes audio and video modalities through separate sub-networks and then integrates these analyses, ensuring alignment with the learned POI model without requiring the POI's training data for testing. This not only sets it apart from existing methods but also offers greater flexibility and broader applicability. The model's robustness against various challenges, such as compression and adversarial attacks, underscores its potential for real-world applications. Notably, experimental results demonstrate that POI-Forensics outperforms other audio-visual and single-modality methods in detection accuracy, representing a significant advancement in the field of deepfake detection.

In the study [5] a groundbreaking model is introduced that harnesses the combined power of audio and visual data to enhance deepfake detection accuracy. AVoiD-DF innovatively addresses the constraints of relying solely on audio or video by implementing a dual-stage architecture, where both modalities are intricately interwoven. The visual aspect utilizes neural networks (CNNs) to capture spatial and temporal characteristics, from video frames while the audio part employs Mel Frequency Cepstral Coefficients (MFCCs) and Gammatone features to conduct a thorough analysis of the audio. A pivotal element of this model is its joint learning mechanism, wherein audio and visual features converge in a unified latent space, enriched by a cross-modal attention mechanism. This feature alignment across modalities significantly boosts the model's proficiency in detecting subtle inconsistencies often missed in single-modality methods. The cross-modal attention, in particular, is crucial as it synergistically merges information from audio and video, forging a comprehensive and accurate data representation. Demonstrating robustness against environmental noise and

camera effects, and delivering top-tier performance on benchmark datasets, AVoiD-DF sets a new standard in deepfake detection, surpassing existing audio-only, video-only, and other joint learning models. The approach adopted by AVoiD-DF marks a substantial leap forward in the field, offering a more integrative and precise solution for identifying deepfakes.

Authors of the [6] propose a novel approach focusing on non-speech audio elements, achieving remarkable accuracy, and demonstrating robustness across various deepfake techniques. This survey reveals the evolving landscape of deepfake detection, highlighting the shift from traditional single-modality methods to more sophisticated, multimodal approaches. The integration of diverse techniques underscores the complexity and urgency of effectively combating deepfake technologies.

The paper [7] tackles the intricate challenge of detecting deepfakes in group conversation settings, addressing the shortcomings of existing methods that struggle in environments with background noise and multiple speakers. Introducing the Group-Aware Deep Convolutional Neural Network (GADCNN), this study innovatively focuses on both individual speaker attributes and group-level interaction dynamics, significantly enhancing detection accuracy and outperforming traditional methods in terms of true positive and false positive rates. Despite its notable success, the paper acknowledges limitations such as the small dataset size and vulnerability to adversarial attacks, suggesting further research with expanded datasets and exploration of countermeasures. The potential integration of GADCNN into real-time conversation systems is also highlighted, pointing towards its practical applicability. Overall, this paper marks a substantial contribution to the field of deepfake detection, particularly in dynamic group settings, paving the way for more sophisticated and robust detection systems.

The paper [8] introduces a machine learning-centric approach to deepfake audio detection, emphasizing the use of Mel-frequency cepstral coefficients (MFCCs). This study focuses on the challenges of audio-only deepfake detection, which is considered more complex than video-based methods. The authors' investigation into MFCC features, known for capturing the spectral characteristics of audio signals, is particularly noteworthy. In their study, the researchers assess the performance of several powerful machine learning algorithms, including Random Forest, Decision Tree, and SVM. Impressively, the SVM classifier proves to be highly effective, achieving an accuracy of over 95% on both real and manipulated audio data from the Fake-or-Real dataset. Further investigation into dimensionality reduction techniques, such as PCA, reveals their advantage in optimizing model accuracy. Despite these promising results, the study is candid in acknowledging its limitations, including a narrow focus on a single dataset and the possible impact of pre-processing methods. Future directions suggested including broader dataset evaluation for enhanced generalizability and the integration of this approach into real-world applications, such as voice assistants or online communication platforms. This research underscores the efficacy of traditional machine learning methods, particularly SVMs, in the realm of audio deepfake detection, providing a viable alternative to more complex deep learning approaches.

This paper [9] introduces a novel approach to deepfake audio detection, leveraging a vision transformer-based methodology that diverges from traditional audio-based techniques. The authors first convert audio signals into spectrograms, transforming the audio frequency content into visual representations.

Subsequently, they employ a vision transformer, originally designed for image analysis, to classify these spectrograms as either genuine or manipulated audio. The results demonstrate that this vision transformer-based approach yields promising performance on a dataset comprising real and deepfake audio samples, showcasing comparable or superior efficacy compared to established audio-based methods. This approach underscores the potential of visual features extracted from spectrograms to capture nuanced manipulation cues not readily discernible in raw audio data. While acknowledging the necessity for broader and more diverse datasets for generalizability, the study suggests exploring pre-training the vision transformer on extensive audio datasets to potentially enhance performance and robustness. Additionally, investigating the combination of this novel approach with traditional audio-based techniques is proposed for the development of a comprehensive and robust deepfake detection system. In summary, this paper introduces an innovative and promising avenue for deepfake audio detection using vision transformers, challenging the conventional audio-centric focus and paving the way for further exploration in this domain.

This paper [10] addresses the multifaceted challenge of detecting deepfakes across different media types, including audio, images, and videos. It categorizes deepfakes into four main types, discussing the limitations of current generation and detection techniques while highlighting the ongoing race between creators and detectors. The authors propose a comprehensive "Deepfake Detection System" model encompassing preprocessing, feature extraction, and classification stages. This paper is a valuable introductory resource for those new to the field, underscoring the need for further research and development. However, it lacks specific implementation details for the proposed model, offers limited comparisons with existing detection methods, and primarily focuses on theoretical aspects. In summary, while providing valuable insights and a framework for future work, this paper could benefit from more in-depth exploration and empirical evaluation of its proposed model and a more extensive review of existing detection approaches.

This paper [11] delves in the realm of deepfakes and how to detect them we're concentrating on harnessing the capabilities of Deep Convolutional Neural Networks (CNNs) to accurately identify audio content. Our approach involves a CNN structure designed specifically for this purpose leveraging the analysis of features such, as Mel Frequency Cepstral Coefficients (MFCCs) and spectral attributes to effectively differentiate between authentic and manipulated audio. The proposed model exhibits promising results on a dataset comprising both genuine and deepfake audio samples, demonstrating high accuracy in identifying manipulated content. The paper highlights the advantages of CNNs in deepfake detection, emphasizing their capacity to learn intricate patterns and maintain robust performance even with limited training data. However, it acknowledges limitations associated with the dataset's size and diversity, suggesting the necessity for further evaluation on larger, more varied datasets to ensure the model's generalizability. Moreover, the authors propose the exploration of various pre-processing methods in order to achieve the most effective feature extraction. They also urge for the investigation of how their proposed model can be integrated into practical applications like online communication platforms and voice assistants. In conclusion, this paper emphasizes the potential of using CNNs as an important tool

for detecting deepfake audio. It presents a promising model architecture and emphasizes the strengths of CNNs in addressing this challenging task. Nevertheless, the authors acknowledge the importance of research and enhancements to guarantee performance, in practical scenarios.

This paper [12] introduces an innovative approach to deepfake audio detection through the utilization of unsupervised pretraining models. It presents two distinct architectures: a feature extraction model and a multi-task learning model, both based on unsupervised pretraining. These models exhibit remarkable performance on the ADD2022 challenge, a benchmark dataset for deepfake audio detection. The feature extraction model achieves an Equal Error Rate (EER) of 32.80% for low-quality fake audio detection, while the multi-task learning model achieves an exceptional 4.80% EER for partially fake audio detection. Notably, the multi-task learning model demonstrates robustness and generalizability, even when trained on substantially different data, making it promising for real-world applications. Although the paper acknowledges potential limitations against high-quality deepfakes, it emphasizes the need for further research to enhance performance in such scenarios. The authors suggest exploring various unsupervised pretraining models and architectures for potential improvements and advocate for investigating the explainability and interpretability of the models' decisions. In general, this paper demonstrates a progress, in identifying audio using unsupervised pretraining models. The remarkable outcomes and possibilities for exploration emphasize the potential of this method in constructing efficient systems, for detecting deepfakes.

This paper [13] introduces an innovative audio anti-spoofing system designed for robustness against deepfakes and spoofing attacks. It capitalizes on the distinctive characteristics of low-frequency sub-band information to ensure reliable detection. The authors' analysis of spectral features in the low-frequency range (below 200 Hz) reveals discernible differences in response to manipulation attempts in real and spoofed audio. Building upon these insights, they devise a feature extraction technique targeting the low-frequency sub-band and a corresponding classifier to Distinguishing between fake audio is a task. The system being suggested here shows capabilities when dealing with a dataset that includes both deepfake audio examples. It exhibits accuracy and resilience, against methods employed to deceive or manipulate the authenticity of audio recordings. Notably, it maintains its effectiveness even in challenging scenarios involving background noise and channel mismatch. Strengths of this work include the introduction of a novel spoofing detection approach that emphasizes low-frequency features, rendering the system less susceptible to noise and channel variations. Nevertheless, limitations encompass the relatively small dataset used, underscoring the need for broader evaluations on larger and more diverse datasets to ensure the system's generalizability. The authors also acknowledge potential vulnerabilities to advanced spoofing techniques that target the low-frequency range, suggesting further research to enhance robustness. Finally, investigating the computational efficiency of the system for real-time applications is recommended. In summary, this paper offers a promising contribution to audio anti-spoofing by harnessing low-frequency sub-band information, presenting an effective method for deepfake and spoofing attack detection, and fostering the development of more resilient audio security systems in the future.

This paper [14] introduces a novel approach to deepfake audio detection, employing bi-level

optimization to enhance robustness against adversarial attacks and manipulation techniques common in deepfake generation. The method comprises two optimization stages: Level 1 involves a deep neural network model assessing audio features and predicting authenticity, while Level 2 employs an adversarial perturbation function to subtly alter the audio signal to deceive the detection model. This iterative process enhances both the detection model and the perturbation function, resulting in a more resilient and adaptive detection system. The proposed bi-level optimization framework outperforms traditional deepfake detection models on a dataset of genuine and manipulated audio samples, achieving superior accuracy and resilience against adversarial attacks. It demonstrates effectiveness in detecting deepfakes employing various manipulation techniques, including voice cloning and speech synthesis. The strengths of this work include its innovative bi-level optimization approach, improved robustness compared to traditional models, and adaptability to adversarial perturbations, rendering it suitable for practical applications. Despite its promising potential, this methodology faces several obstacles. These include the possibility of high computational expenses which may need to be mitigated, as well as vulnerability to complex adversarial attacks aimed at disrupting the optimization process. Additionally, further research in the area of interpretability is necessary in order to fully comprehend how the detection model makes its decisions. Overall, this paper presents a valuable opportunity for deepfake audio detection through bi-level optimization, offering significant advancement in the creation of robust and trustworthy audio verification systems for the future.

This paper [15] addresses the critical challenge of securing voice biometric systems by proposing Quick-SpoofNet, a deep learning model designed for audio deepfake detection within voice anti-spoofing systems. Quick-SpoofNet utilizes innovative techniques, including one-shot learning, metric learning, and spectral feature analysis, to discern subtle differences between real and manipulated audio. The model renders itself strong in its capacity to generate quality deepfakes even with minimal training examples, achieving so through the utilization of both one-shot and metric learning. One of the major strengths of this model lies in its impressive ability to generalize effectively using a small amount of training data, across a range of deepfake generation techniques. This is made possible through the implementation of one-shot and metric learning, allowing for the production of high-quality deepfakes even with limited training samples. Experimental results on a dataset containing genuine and deepfake voice samples showcase Quick-SpoofNet's superior performance, achieving high accuracy and generalizability even to unseen deepfakes. This paper makes a significant contribution by addressing a crucial issue in voice security and offering a promising one-shot learning approach, which holds potential for developing resilient defense systems against evolving audio manipulation threats. However, future research should involve testing on more diverse real-world audio recordings to confirm generalizability, exploring different feature extraction methods, and assessing the model's integration feasibility into existing voice biometric systems.

This paper [16] introduces SpecRNet, an innovative deep learning architecture tailored for efficient audio deepfake detection. SpecRNet employs lightweight convolutional layers and residual blocks to minimize computational demands while maintaining high accuracy. Compared to existing models like LCNN, SpecRNet reduces processing time

by approximately 40%, rendering it suitable for real-time applications, such as online platforms requiring rapid audio content verification. The paper demonstrates SpecRNet's effectiveness across various datasets and under diverse conditions, illustrating its robustness and generalizability. Noteworthy strengths of SpecRNet include its contribution to faster and more accessible deepfake detection, especially in real-time scenarios, and its compatibility with a range of devices, including mobile phones and embedded systems. While the model exhibits good accuracy and robustness, future work should encompass evaluation against real-world deepfakes with complex manipulation techniques, exploration of methods to enhance accuracy and generalizability, and integration with existing audio processing pipelines and security systems for broader adoption. In summary, "SpecRNet'' significantly advances audio deepfake detection by offering a fast, efficient, and accurate model conducive to practical use, contributing to the fight against misinformation and the safeguarding of online communication.

The paper [17] introduces the SE-Res2Net-Conformer architecture, a novel model designed for detecting both synthetic voice and audio splicing. This architecture combines the strengths of SE-Res2Net for local pattern capture and Conformer for global temporal context, effectively extracting essential features from audio signals. The model outperforms previous approaches in synthetic voice detection on the ASVspoof 2019 dataset. Additionally, the paper proposes a new formulation for audio splicing detection, focusing on identifying splicing segment boundaries, making it more amenable to deep learning methods. The strengths of this work include the combination of complementary feature extraction techniques, leading to improved detection accuracy for synthetic voices and spliced audio segments. The model's performance in synthetic voice detection surpasses that of previous approaches. Moreover, the novel formulation of audio splicing detection presents new possibilities for tackling this challenging task with deep learning methods. However, limitations include the use of a limited dataset for evaluation, emphasizing the need for testing on more diverse datasets and real-world audio scenarios. The authors also acknowledge the potential influence of pre-processing techniques on model performance, suggesting exploration of different methods for optimal feature extraction. Investigating the generalizability of the spliced segment detection approach across various splicing techniques and noise conditions is an area for potential future work. In summary, this paper offers a promising approach to audio manipulation detection through the SE-Res2Net-Conformer architecture, presenting improved performance in synthetic voice and spliced audio detection and opening avenues for more robust and effective detection systems in the future.

In the research paper titled [18] the authors propose a novel method for identifying deepfake audio by harnessing the power of Mel-Frequency Cepstral Coefficients and deep learning techniques. This cutting-edge approach presents a promising solution to the growing concern of deepfake audio in the digital landscape. The proposed deep neural network architecture, utilizing MFCC features as input, distinguishes between real and deepfake audio samples. Comparative analysis of various deep learning models reveals that convolutional neural networks (CNNs) outperform others, achieving high accuracy in detecting manipulated audio on a controlled dataset. Finally, this approach utilizes easily accessible MFCC features and deep learning

models to effectively detect deepfakes in a simple and efficient manner. While the model shows considerable accuracy, it is important to note that the study was limited due to a smaller dataset used for evaluation. To address this, further research should be conducted on larger and more diverse datasets to assess the model's generalizability. Additionally, the authors suggest exploring different pre-processing techniques to improve MFCC extraction, as well as examining the model's ability to withstand advanced deepfake generation methods and adversarial attacks. In conclusion, this paper marks a promising beginning in the detection of deepfakes, with potential for further advancement in future studies. By acknowledging its limitations and actively seeking out ways to improve it, we can drive the growth of stronger and more dependable deepfake detection systems in the years to come.

The paper [19] introduces a novel deep learning approach for detecting fake audio messages by employing a hybrid model that combines both recurrent and convolutional neural networks (RNN-CNNs). The RNNs capture temporal dependencies in audio signals, while CNNs extract spatial features from spectrograms, leveraging their complementary strengths. The proposed model demonstrates promising results on a dataset containing real and fake audio messages, achieving high accuracy in identifying manipulated audio. Compared to using only RNNs or CNNs, the combined RNN-CNN approach exhibits superior performance, highlighting the effectiveness of harnessing both temporal and spatial features for deepfake detection. This paper's strengths lie in its innovative approach, combining RNNs and CNNs for potentially more effective feature extraction and classification, with good accuracy observed on a controlled dataset, signifying its potential for practical applications. The focus on RNN-CNNs represents a promising direction for further research and development in deepfake detection, as it combines temporal and spatial analysis to enhance performance. Limitations include the use of a relatively small dataset, necessitating further evaluation on larger and more diverse datasets to confirm generalizability. Additionally, the paper acknowledges the potential impact of pre-processing techniques on model performance, suggesting exploration of different methods for optimal feature extraction and spectrogram generation. Investigating the model's robustness against sophisticated deepfake generation techniques and adversarial attacks presents an avenue for valuable future work. In summary, this paper presents a promising approach for deepfake detection using RNN-CNNs, offering a novel architecture, encouraging results, and the potential for further improvement, thus promoting further research in the field and contributing to the development of more robust and reliable deepfake detection systems in the future.

This paper [20] explores the application of deep learning methods in detecting deepfake audio within the context of digital forensics. Noteworthy findings include the challenge that deepfake audio poses to digital investigations due to its deceptive potential. The paper provides a comprehensive review of existing deepfake audio classification methods and conducts a comparative analysis of various deep learning techniques, encompassing custom architectures and pre-trained models like VGG-16. These methods are evaluated based on their ability to detect deepfakes using extracted audio features, such as MFCC, Mel-spectrum, Chromagram, and spectrograms. The results reveal that custom architectures achieve superior accuracy with Chromagram, Spectrogram, and Mel-Spectrum image

features, while VGG-16 excels with MFCC image features. This research contributes to enhancing the capabilities of forensic investigators in distinguishing between real and synthetic voices. The strengths of this work include its comprehensive overview of deepfake audio, the evaluation and comparison of various deep learning techniques for forensic applications, the visualization of different audio features to highlight distinctions between real and fake audio, and its provision of valuable insights for advancing digital forensics tools and methods. However, limitations encompass the use of a limited dataset, warranting further assessment with larger and more diverse datasets, a focus on specific deep learning architectures and feature sets, suggesting exploration of other combinations for potential improvements, and the necessity of investigating the models' robustness against advanced deepfake techniques and adversarial attacks. To put it briefly this paper greatly enhances our understanding of identifying audio in forensics. By evaluating and comparing deep learning approaches it provides knowledge, for improving the tools and techniques used to analyze and verify audio evidence, in forensic investigations.

Author of the paper [21] introduces a novel approach to deepfake detection by focusing on the simultaneous analysis of both audio and video modalities. The motivation behind this approach arises from the vulnerability of existing single-modality detection methods to manipulations targeting the other modality. The proposed deep learning architecture enables cross-modal interaction and information fusion, allowing the model to learn relationships and inconsistencies between audio and video features, potentially enhancing detection accuracy. The evaluation on various deepfake datasets demonstrates the effectiveness of this approach in detecting both audio and video deepfakes, even when combined. The strengths of this work include its multimodal approach, which addresses the limitations of single-modality detection, and its robustness, as indicated by good performance on diverse deepfake datasets. However, the paper acknowledges the limitations of existing datasets for multimodal deepfake detection, emphasizing the need for larger and more diverse datasets for further evaluation. Additionally, it highlights the importance of improving explainability and understanding the model's decision-making process, given the potential reduced interpretability of multimodal models. Further research is required to assess the model's performance in real-world scenarios involving potentially more sophisticated deepfakes and adversarial attacks. In summary, this paper offers a promising approach for robust deepfake detection through the simultaneous analysis of audio and video modalities, addressing the limitations of single-modal methods and potentially leading to more reliable and effective deepfake detection systems in the future.

This paper [22] critically examines the potential impact of deepfakes on the dissemination of scientific knowledge and proposes strategies to mitigate their adverse effects. Findings indicate that individuals in the education sector, including adults and educators, are susceptible to deepfake manipulation, given their trust in information sources and reliance on video content for learning and communication. The study underscores the significance of developing tools and strategies for detecting deepfakes, promoting critical thinking skills, and encouraging information verification before acceptance. A field experiment is conducted to assess vulnerability, with participants exposed to both authentic and manipulated science videos. The results highlight the need for

targeted interventions and educational training, particularly for adults and educators. The study's strengths lie in its exploration of a relatively under-researched area, emphasizing the multi-pronged approach required to combat deepfakes, including technological solutions, educational initiatives, and social awareness campaigns. Despite its contributions, this study has limitations due to its small sample size. Therefore, further research involving larger and more diverse populations is necessary. Additionally, it is crucial to investigate the effectiveness of different detection and mitigation strategies in real-world situations. Moreover, there is potential for future research to explore the use of emerging technologies such as blockchain and digital fingerprinting to ensure tamper-proof dissemination of knowledge. In conclusion, this paper serves as a timely and valuable addition to the discourse surrounding deepfakes and their potential impact on the reliability and integrity of scientific information dissemination. It highlights the importance of being proactive and taking measures to safeguard the dissemination of scientific knowledge.

3. Proposed Architecture

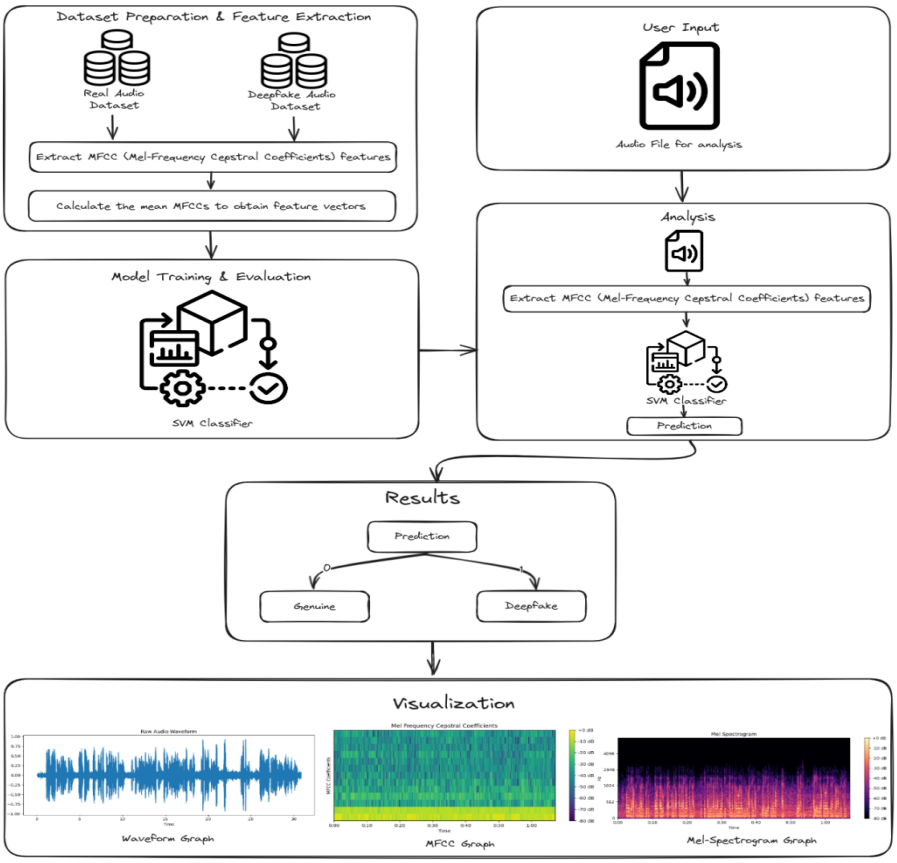

Figure 1: Architecture of the Proposed Model

Figure 1 explains the architecture our system, which begins with the preparation of the datasets. We curate a collection of genuine audio clips and an equivalent set of sophisticated deepfake audio samples. The authenticity of real audio samples is verified through controlled recording environments to ensure the baseline dataset's integrity. We utilize Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction due to their

effectiveness in encoding timbral aspects of the audio signal which are crucial for distinguishing deepfakes from real audio. To streamline the dataset, we compute the mean MFCCs across all samples to derive a consistent feature vector that represents the essence of the dataset. This process ensures a reduced-dimensional feature space for efficient training.

Our predictive model, as displayed in figure 1 of the model training and evaluation section, relies on a customized Support Vector Machine (SVM) classifier. We specifically selected an SVM due to its impressive performance in high-dimensional spaces and its capacity to handle non-linear boundaries through kernel functions. To ensure the model's effectiveness in predicting unseen samples, we conduct a grid search optimization to fine-tune hyperparameters. Additionally, we employ cross-validation with a subset of the dataset that was not used in the training process, using metrics like accuracy, precision, recall, and F1-score to continuously enhance the model's performance.

The user interface accepts an audio file input, which is then processed to extract MFCCs, mirroring the feature extraction process used in dataset preparation this is shown in figure 1 in user input. These features are fed into the SVM classifier, which uses the decision function shaped during the training phase to evaluate the audio file. The classifier outputs the a score that provides an indication of the probability that the audio's a deepfake. To ensure robustness, we implement a thresholding mechanism that allows for configurable sensitivity, accommodating scenarios where a higher degree of certainty is required before flagging an audio clip as fake.

The prediction made by the SVM classifier is presented to the user along with a confidence score that quantifies the certainty of the model's decision which is shown in figure 1 in results. Visualization tools are shown in figure 1 in visualization, are integrated into the system to offer a transparent view of the decision-making process: The MFCC graph visually represents the extracted features from the user's audio file, allowing for a comparison against typical profiles of real and fake audio. The waveform graph provides a direct visual comparison of the audio file's waveform to common patterns observed in genuine and deepfake samples. The Mel-spectrogram offers a heat map of frequency intensities over time, providing insight into the temporal characteristics of the audio signal, which could be indicative of manipulation. These visual outputs not only serve as an explanatory aid to support the system's prediction but also enable users to perform a heuristic analysis, potentially identifying artifacts that automated processes may overlook.

4. Implementation

4.1 Dataset Preparation

For the construction and evaluation of our SVM-based deepfake audio detection system, we employed the 'for-original' variant of the Fake-or-Real (FoR) Dataset. This dataset is part of a comprehensive collection curated by the APTLY lab and accessible through the Biometric Intelligence Lab at York University [23]. The 'for-original' dataset comprises a substantial corpus of over 195,000 audio utterances, meticulously gathered to represent both authentic human speech and synthetic speech outputs from state-of-the-art TTS technologies. Our system's design philosophy mandated the use of raw, unaltered data to ensure that the model was trained under conditions that closely mimic real-world scenarios. This dataset variant, being the most pristine and unprocessed among the available options, was thus an ideal fit for our objectives.

Dataset Characteristics:

Volume and Diversity: The "for-original" dataset contains a diverse collection of speech variations that encompass a broad range of vocal characteristics shaped by the speaker's identity, accent, and linguistic content.

Source Inclusivity: The inclusion of samples from advanced TTS systems like Deep Voice 3 and Google Wavenet TTS, alongside human speech from the Arctic, LJSpeech, and VoxForge datasets, provides a robust challenge for the classifier's discriminatory capacity.

Quality Assurance: The high fidelity of the recordings ensures that the model is trained and tested against data that maintain the integrity of the acoustic properties inherent in genuine and synthetic speech

4.2 Feature Extraction

MFCCs are widely recognized for their efficacy in encoding timbral and textural aspects of sound, making them particularly suitable for speech and audio analysis tasks where the identification of unique characteristics is paramount. The process of computing MFCCs entails several computational stages, each designed to transform the raw audio waveform into a feature set that faithfully captures the essential spectral properties while aligning with the human auditory system's perceptive capabilities.

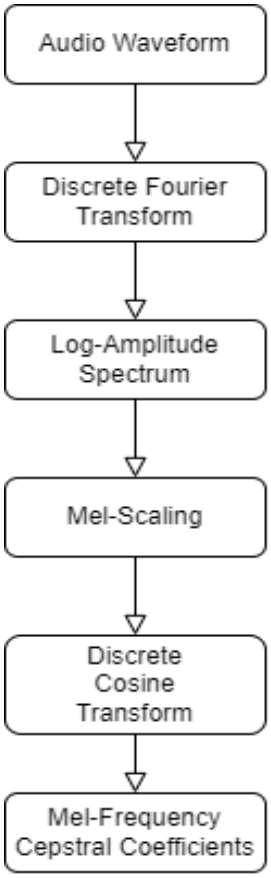

Figure 2: Process of extraction of MFCCs

Process of Computing MFCCs is explained in figure 2. The process begins with the raw audio waveform $x(t)$, representing the sound pressure variations over time. The first computational step is the application of the DFT, which transforms the signal from the time domain into the frequency domain. The DFT of an audio sample is mathematically represented as:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\frac{2\pi}{N}kn} \quad \textbf{(1)}$$

where $X(k)$ is the Fourier Transform of the signal at frequency bin k, $x(n)$ is the **n-th** sample of the input signal, N and is the total number of samples. The magnitude squared of the DFT results in the power spectrum, which illustrates the power present at each frequency component:

$$P(k) = |X(k)|^2$$

The power spectrum is then passed through a set of bandpass filters known as the Mel filter bank. The number of filters, *M*, in the filter bank typically ranges from 20 to 40 and is spaced uniformly on the Mel scale which is shown in figure 3. The filter bank output, *S*(*m*), is given by:

$$S(m) = \sum_{k=0}^{N-1} P(k) \cdot H_m(k)$$
(3)

where $H_m(k)$ is the Mel filter bank's ***m*-th** filter.

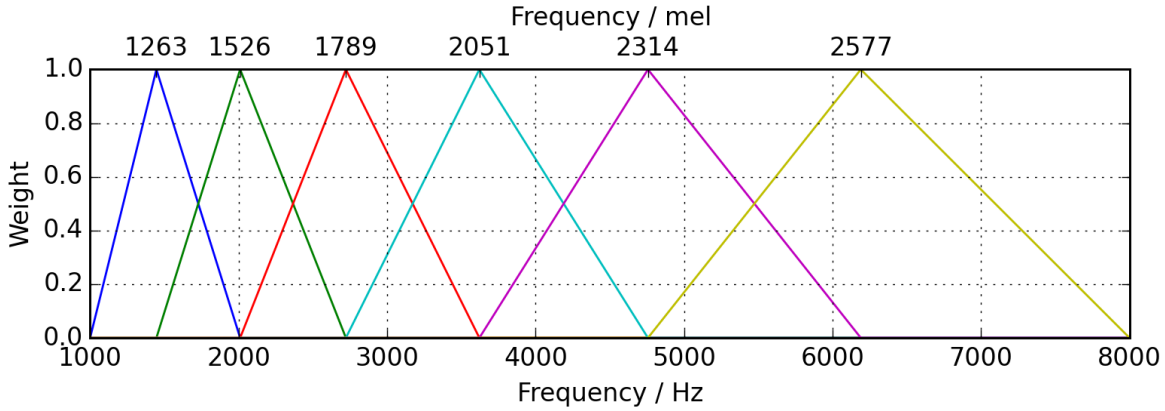

Figure 3: Mel Scale

The log filter bank energies are calculated using a logarithmic scale, mimicking the way our ears perceive loudness, and producing a group of precise measurements:

$$log\ S(m) = lo\left(\sum_{k=0}^{N-1} P(k) \cdot H_m(k)\right)$$
(4)

Finally we apply the Discrete Cosine Transform (DCT) to the log Mel filter bank energies to calculate the MFCCs. This step decorrelates the log Mel spectrum and yields a compressed representation of the filter banks, emphasizing the lower order coefficients, which typically capture the most salient aspects of the signal. The ***n*-th** MFCC, ***Cn***, is calculated as follows:

$$C_n = \sum_{m=1}^{M} log\ log\ S\ (m) \cdot cos\ cos\ \left[n(m - 0.5)\frac{\pi}{M}\right]$$
(5)

for ***n*=1,2,..., *L***, where ***L*** is the number of MFCCs kept for the analysis (often ***L*** is set to **12 or 13**).

The MFCCs are a compact representation of the audio signal's spectral characteristics. The lower-order coefficients, which contain the most important information for audio processing tasks, are typically utilized for deepfake detection. This selection is due to their ability to characterize the vocal tract configuration, which is altered during the creation of deepfake audio. In deepfake detection algorithms, these coefficients serve as input features to classification models, such as Support Vector Machines (SVM). Their effectiveness stems from their capacity to capture nuances in speech that can distinguish genuine from manipulated audio. The MFCCs' robustness against variations in speaking environments and recording conditions further justifies their selection for this application.

For our system's core analytical capability hinges on extracting Mel-Frequency Cepstral Coefficients (MFCCs) to serve as features for our Support Vector Machine (SVM) classifier.

The `extract_mfcc_features` function processes audio files to compute 13 MFCCs, utilizing an FFT window of 2048 and a hop length of 512. These parameters were empirically determined to capture the essential characteristics of the audio signal for the purpose of deepfake detection. The dataset is dynamically constructed using the `create_dataset` function, which iterates over audio files in specified directories, classifying them as genuine or deepfake. The function extracts MFCC features from each audio sample and labels them accordingly, ensuring a balanced dataset for model training.

4.3 Model Training & Evaluation

Before we train our SVM classifier, it's essential to preprocess our feature set by standardizing it with a mean of zero and a variance of one. To accomplish this, we rely on the `StandardScaler` from Scikit-learn. Our training process involves splitting the dataset into two sets, a training set and a test set, using a stratified approach to maintain the proportion of classes between them. Using Scikit-learn's `SVC` with a linear kernel, we then train our SVM classifier on the scaled training data.

Post-training, the classifier's performance is quantified through the accuracy metric, and the results are distilled into a confusion matrix. These metrics play a crucial role in evaluating how well the classifier can apply what it learned from the training data to new and unseen data, giving an unbiased indication of its predictive capabilities. We deliberately chose accuracy as our primary metric due to its interpretability and relevance to binary classification problems; however, it's complemented by the confusion matrix, which provides deeper insight into classification errors. To facilitate the operational deployment of the model, we serialize the trained SVM classifier and the scaler using Joblib, which is a Python library for lightweight pipelining in Python. This allows for the model and preprocessing steps to be saved and loaded efficiently for subsequent predictive analysis without the need to retrain. The technical architecture of our model training pipeline is designed to ensure scalability, performance, and maintainability. This approach allows us to adapt our solution to the evolving landscape of deepfake audio detection, ensuring that our system remains at the forefront of technological efficacy.

The evaluation of the SVM model has been conducted on a dataset comprising real and deepfake audio samples.

```
Unique classes in y_train: [0 1]
Size of X: (30204, 13)
Size of y: (30204,)
Size of X_train: (24163, 13)
Size of X_test: (6041, 13)
Size of y_train: (24163,)
Size of y_test: (6041,)
Accuracy: 0.9728521767919218
Confusion Matrix:
[[5309   79]
 [  85  568]]
```

Figure 4: Model Evaluation result

The performance metrics extracted described in figure 4 from the testing phase are as follows:

The dataset is composed of two distinct classes - 0 for genuine and 1 for deepfake. These classes were identified through careful examination of the unique classes within the training set, which consists of a total of 30,204 samples. Each sample is described by 13 MFCC features, indicating a sizable dataset. This large dataset size is advantageous for building a robust model. To enable effective training and evaluation, the dataset was divided into a training set comprising 24,163 samples and a test set with 6,041 samples, adhering to the standard 80-20 split ratio. This common practice in machine learning ensures sufficient

data for learning while also providing a substantial evaluation set.

Figure 5 displays the confusion matrix for the test set.

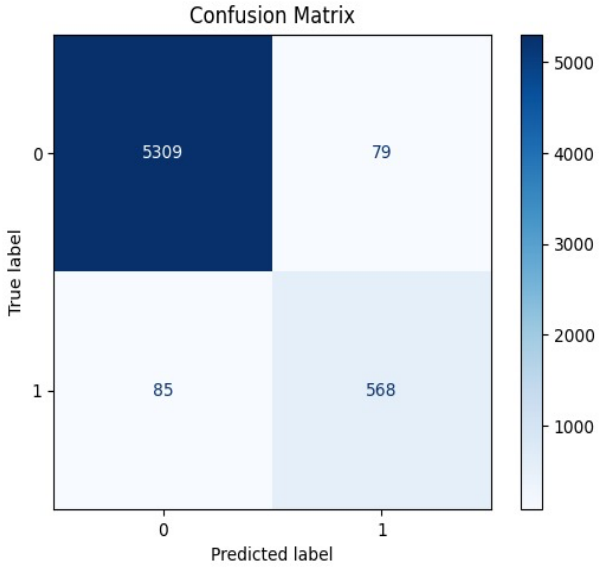

Figure 5: Confusion Matrix for the model

Where TP (True Positive) indicates genuine audio correctly classified, FP (False Positive) indicates genuine audio incorrectly classified as deepfake, FN (False Negative) indicates deepfake audio incorrectly classified as genuine, and TN (True Negative) indicates deepfake audio correctly classified.

The confusion matrix provides insights:

- The model performed significantly well, having a minimal false positive rate of only 79 out of 6469 genuine samples incorrectly classified. This is particularly crucial for situations where falsely identifying authentic audio as deepfake could have severe consequences.

- The model exhibits a low false negative rate, with only 85 out of 647 deepfake samples being incorrectly labeled. This further demonstrates the model's reliability in successfully identifying the majority of deepfake instances, a crucial component in the success of deepfake detection systems.

5. Results & Visualization

5.1 Results

The model's accuracy of 97.28% is a strong indication of its ability to effectively differentiate between real and deepfake audio samples, as illustrated in figure 4. Such high success rate serves as a testament to the model's proficiency.

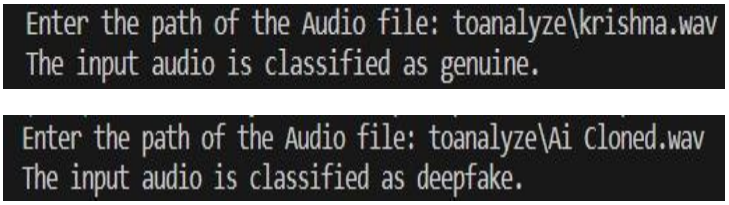

Figure 6: Result of input audio

The result shown in figure 6 is generated at the end of testing the various audio files as shown in the image.

The SVM classifier showed exceptional performance, achieving a remarkable 97.28% classification accuracy in accurately distinguishing genuine and deepfake audio samples. This impressive outcome highlights the strong predictive ability of the model within the specific parameters of the test setting. The model's high true positive and true negative rates further demonstrate its proficiency in confidently identifying both classes. Moreover, the balance observed in the representation of both classes during the training and testing phases serves as a testament to the robustness of the SVM classifier. This equilibrium ensures that the model does not exhibit any bias towards a particular class. However it's crucial to be cautious when interpreting these findings in real life situations because variables, like quality, background noise and recording circumstances have the potential to impact the systems reliability.

5.2 Visualization

Waveform Plot

This part of the study emphasizes the visualization of audio waveforms, providing a comparative analysis between real and deepfake audio samples. These visualizations facilitate an understanding of the variations inherent in genuine versus manipulated audio content.

Real Audio Waveform:

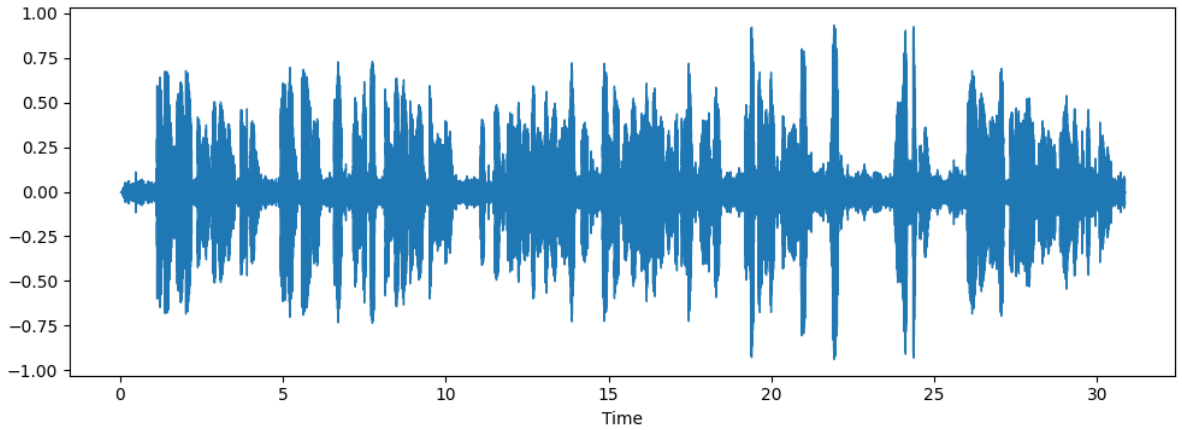

Figure 7: Raw Real Audio Waveform

The waveform in Figure 7 represents a genuine audio sample titled real_audio.wav.

Deepfake Audio Waveform:

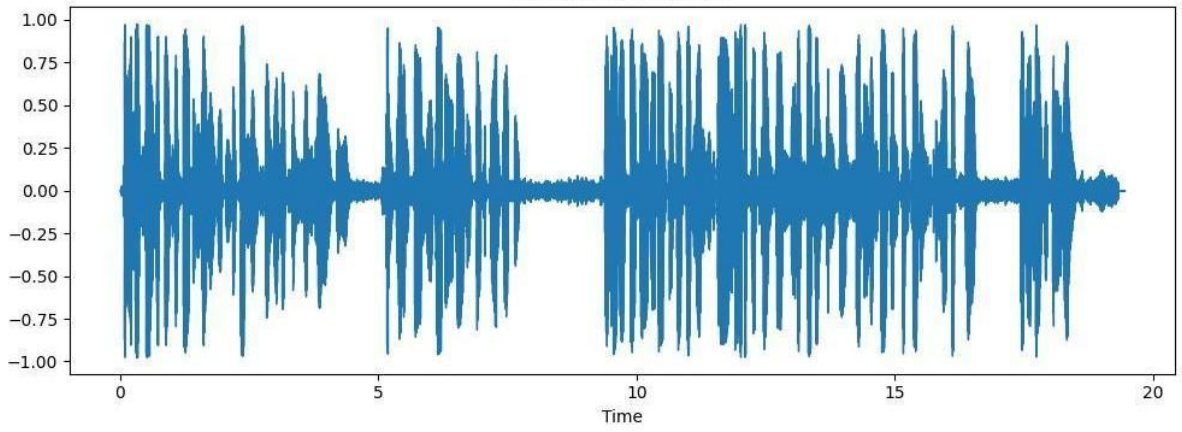

Figure 8: Deepfake Audio Waveform

Figure 8 illustrates the waveform of a deepfake audio sample.

Spectrogram

In this study, we incorporate spectrogram analysis as a crucial component to enhance our deepfake audio detection methodology. Spectrograms, with their ability to visually display the frequency spectrum of audio signals over time, offer indispensable insights into the complex interplay of frequencies in both authentic and manipulated audio. This analytical approach is essential for identifying subtle spectral anomalies that are characteristic of deepfake audio, thereby providing a robust tool for our comparative analysis.

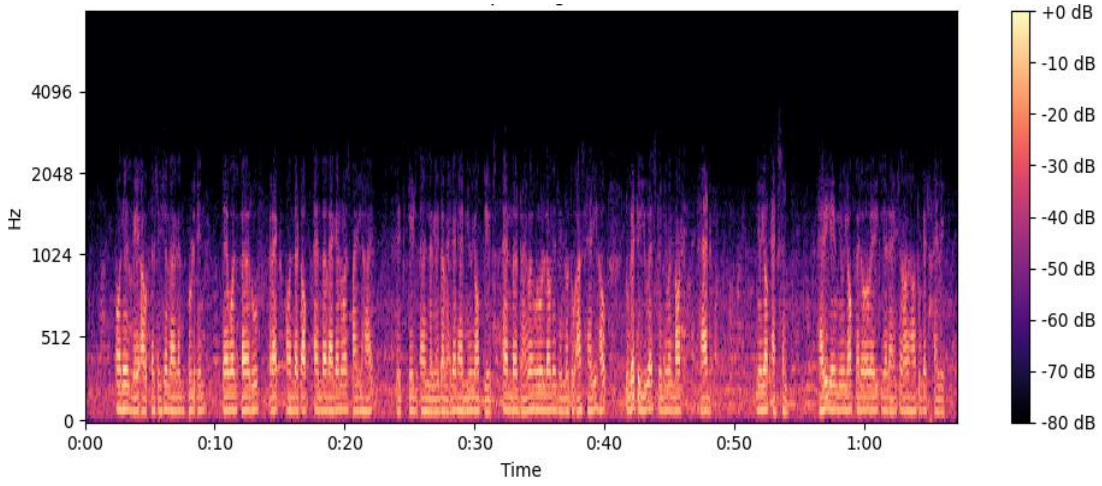

Figure 9: Mel Spectogram of Real Audio

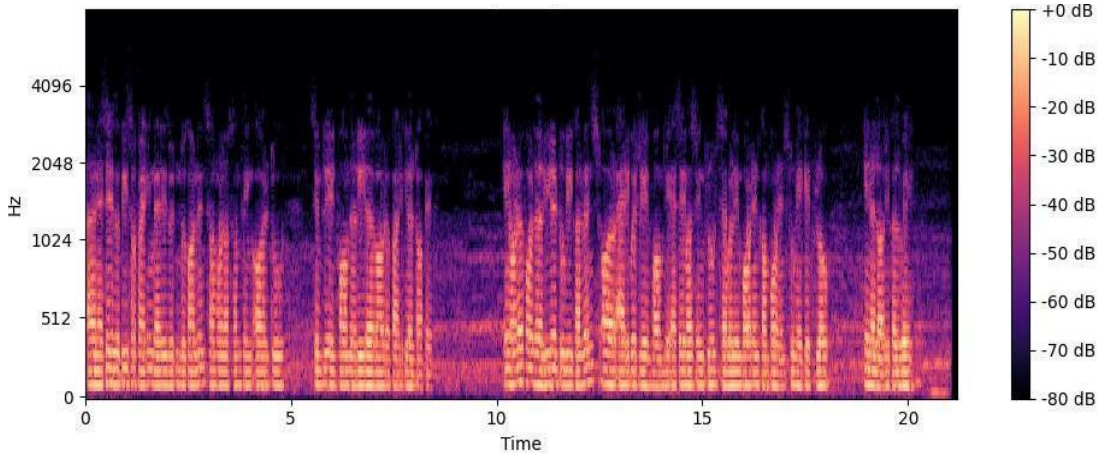

Figure 10: Mel Spectogram of Deepfake Audio

In Figure 9 & 10, we present two Mel spectrogram images: the first depicting real audio with its natural, fluctuating frequency patterns, and the second showing deepfake audio, characterized by irregular spectral features. These visual contrasts, especially in the spectral energy distribution, are critical in differentiating authentic speech from synthetically generated content.

MFCC Plot

MFCCs, also known as Mel-Frequency Cepstral Coefficients, hold immense importance in the realm of audio signal processing, specifically in speech and audio recognition. Their efficiency lies in their ability to encapsulate the power spectrum of an audio signal in a condensed form. This allows for the extraction of crucial timbral features, enabling the differentiation of various sounds and voices. In the realm of deepfake detection, MFCCs play a critical role in identifying minute changes and modifications in speech patterns, which are common in manipulated audio.

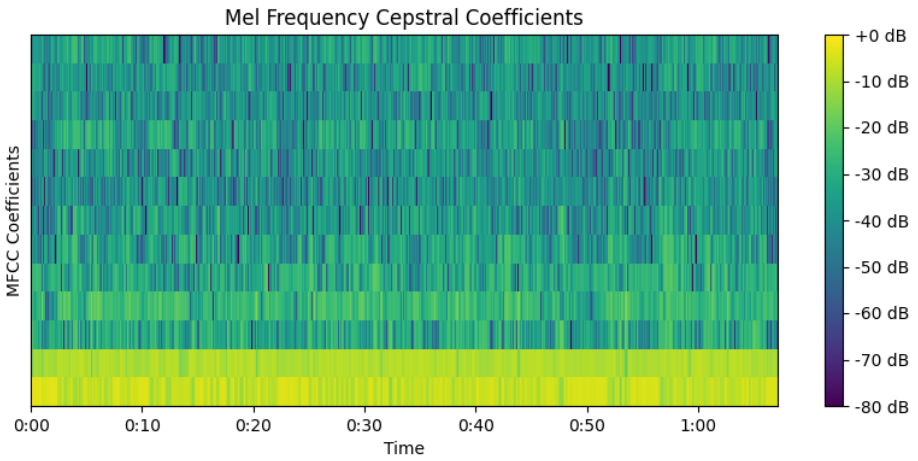

Figure 11: MFCC graph for Real Audio

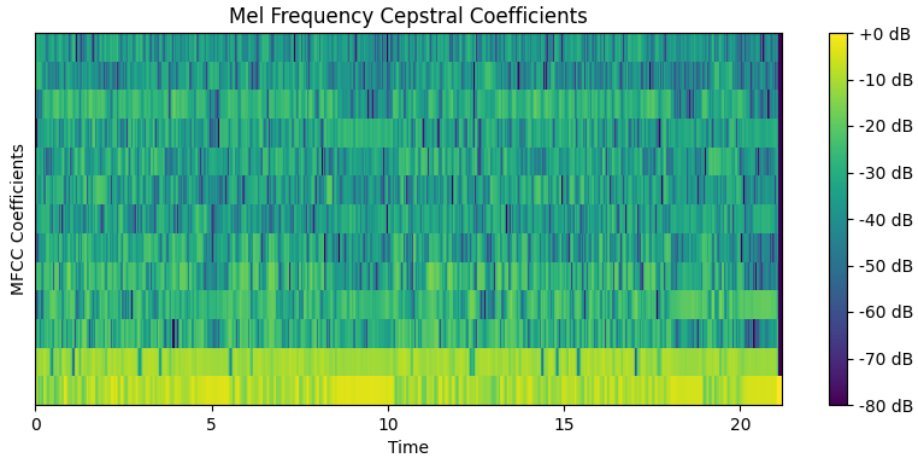

Figure 12: MFCC graph for Deepfake Audio

The comparison between the MFCC plots of the authentic and deepfake audio reveals distinct differences which are shown in figure 11 & figure 12. The authentic audio's MFCC plot shows a consistent and regular pattern of cepstral features, aligning with the expected characteristics of natural speech. In contrast, the deepfake audio's plot displays irregularities and inconsistencies in the cepstral coefficients, indicating potential manipulation.

6. Conclusion

This research presents a significant advancement in the detection of synthetic audio, addressing a critical challenge in the era of rapidly evolving deepfake technology. By employing a Support Vector Machine (SVM) classifier, trained on the comprehensive 'for-original' Fake-or-Real (FoR) dataset, our study demonstrates a sophisticated approach to distinguishing genuine speech from its synthetic counterparts. The successful extraction and utilization of Mel-Frequency Cepstral Coefficients (MFCCs) as features have culminated in the model achieving an impressive 97.28% accuracy, underscoring its potential as a robust tool in the realm of digital authentication and security. The strategic selection of the FoR dataset, notable for its diversity and volume, provided a realistic and challenging testing ground for our model. Not only did this method improve the model's performance in test environments, but it also demonstrated its practicality in real-life

scenarios where the legitimacy of digital audio is often challenged.

In the broader context, our work contributes significantly to the fields of cybersecurity and digital forensics, where the ability to accurately detect deepfake audio is of paramount importance. As deepfake technology becomes more sophisticated, tools such as ours are essential in the fight against digital fraud and misinformation. Looking forward, the research paves the way for further exploration and refinement. Future initiatives will focus on expanding the dataset with more diverse synthetic speech types, experimenting with advanced machine learning algorithms, and improving the interpretability and user interface of the system. These enhancements will aim to not only improve the precision and dependability of the detection mechanism but also to make it more accessible and usable for a wider range of applications. In conclusion, this research marks a pivotal step in the development of effective deepfake detection methods. It lays a strong foundation for ongoing efforts to safeguard digital information integrity in an era where distinguishing between real and artificial has never been more crucial.

References

[1] M. A. Khder, S. Shorman, D. T. Aldoseri, and M. M. Saeed, “Artificial Intelligence into Multimedia Deepfakes Creation and Detection,” in 2023 International Conference on IT Innovation and Knowledge Discovery, ITIKD 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ITIKD56332.2023.10099744.

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

[3] H. H. Kilinc and F. Kaledibi, “Audio Deepfake Detection by using Machine and Deep Learning,” in Proceedings - 10th International Conference on Wireless Networks and Mobile Communications, WINCOM 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/WINCOM59760.2023.10323004.

[4] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, “Audio-Visual Person-of-Interest DeepFake Detection,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, 2023, pp. 943–952. doi: 10.1109/CVPRW59228.2023.00101.

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

[6] T. P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, “BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICASSP49357.2023.10095927.

[7] R. L. M. A. P. C. Wijethunga, D. M. K. Matheesha, A. Al Noman, K. H. V. T. A. De Silva, M. Tissera, and L. Rupasinghe, “Deepfake audio detection: A deep learning based solution for group conversations,” in ICAC 2020 - 2nd International Conference on Advancements in Computing, Proceedings, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 192–197. doi: 10.1109/ICAC51239.2020.9357161.

[8] A. Hamza et al., "Deepfake Audio Detection via MFCC features using Machine Learning," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3231480.

[9] G. Ulutas, G. Tahaoglu, and B. Ustubioglu, "Deepfake audio detection with vision transformer based method," in 2023 46th International Conference on Telecommunications and Signal Processing, TSP 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 244–247. doi: 10.1109/TSP59544.2023.10197715.

[10] B. F. Nasar, T. Sajini, and E. R. Lason, "Deepfake Detection in Media Files - Audios, Images and Videos," in 2020 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 74–79. doi: 10.1109/RAICS51191.2020.9332516.

[11] B. Kumar and S. R. Alraisi, "Deepfakes Audio Detection Techniques Using Deep Convolutional Neural Network," in 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 463–468. doi: 10.1109/COM-IT-CON54601.2022.9850771.

[12] Z. Lv, S. Zhang, K. Tang, and P. Hu, "FAKE AUDIO DETECTION BASED ON UNSUPERVISED PRETRAINING MODELS," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 9231–9235. doi: 10.1109/ICASSP43922.2022.9747605.

[13] M. Li and X. P. Zhang, "Robust Audio Anti-Spoofing System Based on Low-Frequency Sub-Band Information," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/WASPAA58266.2023.10248132.

[14] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Robust Deepfake Audio Detection via Bi-Level Optimization," Institute of Electrical and Electronics Engineers (IEEE), Dec. 2023, pp. 1–6. doi: 10.1109/mmsp59012.2023.10337724.

[15] A. Khan and K. M. Malik, "Securing Voice Biometrics: One-Shot Learning Approach for Audio Deepfake Detection," in 2023 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/WIFS58808.2023.10374968.

[16] P. Kawa, M. Plata, and P. Syga, "SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection," in Proceedings - 2022 IEEE 21st International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 792–799. doi: 10.1109/TrustCom56396.2022.00111.

[17] L. Wang, B. Yeoh, and J. W. Ng, "Synthetic Voice Detection and Audio Splicing Detection using SE-Res2Net-Conformer Architecture," in 2022 13th International Symposium on Chinese Spoken Language Processing, ISCSLP 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 115–119. doi: 10.1109/ISCSLP57327.2022.10037999.

[18] I. Altalahin, S. Alzu'Bi, A. Alqudah, and A. Mughaid, "Unmasking the Truth: A Deep Learning Approach to Detecting Deepfake Audio Through MFCC Features," in 2023 International

Conference on Information Technology: Cybersecurity Challenges for Sustainable Cities, ICIT 2023 - Proceeding, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 511–518. doi: 10.1109/ICIT58056.2023.10226172.

[19] A. Khovrat and V. Kobziev, "Using Recurrent and Convulation Neural Networks to Indentify the Fake Audio Messages," in 2023 IEEE 7th International Conference on Methods and Systems of Navigation and Motion Control, MSNMC 2023 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 174–177. doi: 10.1109/MSNMC61017.2023.10329236.

[20] M. McUba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," in Procedia Computer Science, Elsevier B.V., 2023, pp. 211–219. doi: 10.1016/j.procs.2023.01.283.

[21] D. Salvi et al., "A Robust Approach to Multimodal Deepfake Detection," J Imaging, vol. 9, no. 6, Jun. 2023, doi: 10.3390/jimaging9060122.

[22] C. Doss et al., "Deepfakes and scientific knowledge dissemination," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-39944-3.

[23] Members of APTLY lab, "Fake-or-Real Audio Dataset." Accessed: Jan. 20, 2024. [Online]. Available: https://www.eecs.yorku.ca/~bil/Datasets/for-original.tar.gz