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

Improving Deepfake Audio Detection: A Support Vector Machine Approach with Mel-Frequency Cepstral Coefficients

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

Submitted: 27/12/2023 **Revised**: 03/02/2024 **Accepted**: 11/02/2024

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: *Audio Analysis, Deepfake Detection, Feature Extraction, Media Manipulation, Mel-Frequency Cepstral Coefficients (MFCCs), Support Vector Machine (SVM),*

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

[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 : 0009-0005-4065-4361
[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 : 0009-0003-1266-3709
[8] *Student, 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 : 0009-0000-9677-9175
** Corresponding Author Email: shwetambari.borade@sakec.ac.in*

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

Authors in [1] provide a comprehensive overview of deepfake creation and detection, focusing on multimedia content. They explore various generation techniques but lack detailed insights into specific detection models or their performance. The paper emphasizes the need for robust detection methods and highlights challenges in deepfake technology. However, it lacks an in-depth analysis or concrete results on audio-based detection methods, including MFCCs and SVMs, which are central to our research.

In [2], the authors extensively explore various approaches to create audio deepfakes. The paper categorizes methods, including voice conversion, speech synthesis, voice cloning, and audio editing. It highlights advancements in voice conversion and speech synthesis, making fake audio more realistic. Voice cloning is recognized as a challenging area requiring extensive training data. The study also notes the effectiveness of audio editing but emphasizes its potential for inconsistencies and detectability issues due to unnatural transitions, providing a nuanced understanding of the complexities in audio deepfake creation.

In [3], the paper explores machine and deep learning models for detecting deepfakes, comparing Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and Convolutional Neural Networks (CNNs). The study emphasizes the superiority of ANNs and CNNs over SVMs in audio deepfake detection, with a focus on Mel-Frequency Cepstral Coefficients (MFCCs) as effective features. The findings highlight the efficacy of hybrid approaches, suggesting that combining various models and features can optimize deepfake detection performance.

In [4], 'POI-Forensics' is introduced as a novel approach to deepfake detection, broadening traditional methods with combined audio-visual analysis. The methodology utilizes contrastive learning to distinguish genuine and manipulated representations of specific individuals ('Persons-of-Interest' or POI). Unique in its approach, POI-Forensics processes audio and video separately, integrating analyses through separate sub-networks. This approach ensures alignment with the learned POI model without requiring the POI's training data for testing, offering flexibility and broader applicability. The model demonstrates robustness against challenges like compression and adversarial attacks, outperforming other audio-visual and single-modality methods in detection accuracy and representing a significant advancement in deepfake detection.

In [5], AVoiD-DF presents a groundbreaking model that enhances deepfake detection accuracy by combining audio and visual data. It adopts a dual-stage architecture, intricately weaving neural networks (CNNs) for spatial and temporal characteristics in the visual aspect and Mel Frequency Cepstral Coefficients (MFCCs) and Gammatone features for audio analysis. The model's joint learning mechanism converges audio and visual features in a unified latent space, enriched by a cross-modal attention mechanism. This alignment significantly improves the model's proficiency in detecting subtle inconsistencies often missed in single-modality methods. Demonstrating robustness against environmental noise and camera effects, AVoiD-DF outperforms existing audio-only, video-only, and other joint learning models, setting a new standard in deepfake detection. The approach marks a substantial leap forward, offering a more integrative and precise solution for identifying deepfakes.

In [6], authors propose a novel approach emphasizing non-speech audio elements, achieving remarkable accuracy and robustness across diverse deepfake techniques. The survey showcases the evolving landscape of deepfake detection, emphasizing a shift from traditional single-modality to sophisticated multimodal methods. The integration of diverse techniques underscores the complexity and urgency of effectively combating deepfake technologies.

In [7], the paper addresses the intricate challenge of detecting deepfakes in group conversations, overcoming existing method shortcomings in noisy environments. The Group-Aware Deep Convolutional Neural Network (GADCNN) focuses on individual speaker attributes and group-level dynamics, significantly enhancing detection accuracy and outperforming traditional methods in true positive and false positive rates. Despite notable success, the paper acknowledges limitations such as a 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 highlighted, emphasizing its practical applicability. Overall, the paper contributes substantially to deepfake detection in dynamic group settings, paving the way for more sophisticated and robust detection systems.

In [8], the paper introduces a machine learning-focused approach for audio deepfake detection, emphasizing Mel-frequency cepstral coefficients (MFCCs). The study

addresses challenges in audio-only deepfake detection, showcasing the effectiveness of SVMs with an accuracy exceeding 95% on real and manipulated audio data from the Fake-or-Real dataset. The researchers also explore dimensionality reduction techniques like PCA, enhancing model accuracy. While acknowledging limitations like a narrow dataset focus and potential pre-processing impacts, the study suggests future directions, including broader dataset evaluations for improved generalizability and integration into real-world applications. This research highlights the efficacy of traditional machine learning methods, especially SVMs, in audio deepfake detection, offering a practical alternative to more complex deep learning approaches.

In [9], a novel approach to deepfake audio detection is introduced, utilizing a vision transformer-based methodology distinct from traditional audio techniques. The authors convert audio signals into spectrograms and employ a vision transformer for classification, showcasing promising performance on a dataset with real and deepfake audio samples. This approach highlights the potential of visual features from spectrograms to capture nuanced manipulation cues not easily discernible in raw audio data. While recognizing the need for broader datasets, the study suggests pre-training the vision transformer on extensive audio datasets to enhance performance and robustness. The paper proposes investigating the combination of this approach with traditional audio-based techniques for a comprehensive deepfake detection system, offering an innovative avenue in this domain.

In [10], the paper addresses the challenge of detecting deepfakes across various media types. It categorizes deepfakes, discusses limitations, and proposes a "Deepfake Detection System" model. However, it lacks specific implementation details, comprehensive comparisons with existing methods, and empirical evaluations. While valuable for newcomers, the paper primarily focuses on theoretical aspects, emphasizing the need for further research and development. It provides insights and a framework for future work but could benefit from more in-depth exploration and evaluation of its proposed model and a more extensive review of existing detection approaches.

In [11], the paper focuses on detecting deepfake audio using a specialized Deep Convolutional Neural Network (CNN) structure. The approach utilizes features like Mel Frequency Cepstral Coefficients (MFCCs) and spectral attributes to differentiate between genuine and manipulated audio, showing high accuracy on a dataset with both types of samples. The paper underscores the advantages of CNNs in deepfake detection, citing their ability to learn intricate patterns and maintain robust performance with limited training data. Acknowledging dataset limitations, the authors recommend further evaluation on larger, more diverse datasets for generalizability. They propose exploring different pre-processing methods for effective feature extraction and suggest investigating practical applications like online communication platforms and voice assistants. In conclusion, the paper highlights CNNs as a promising tool for deepfake audio detection, providing a strong model architecture while emphasizing the need for ongoing research and enhancements for practical scenarios.

In [12], the paper introduces an innovative approach to deepfake audio detection using unsupervised pretraining models. It presents two architectures: a feature extraction model and a multi-task learning model, both achieving remarkable performance on the ADD2022 challenge, a benchmark dataset. The feature extraction model achieves a 32.80% Equal Error Rate (EER) for low-quality fake audio, while the multi-task learning model achieves an exceptional 4.80% EER for partially fake audio. The latter demonstrates robustness and generalizability, even with substantially different data, suggesting real-world applicability. Acknowledging potential limitations against high-quality deepfakes, the paper calls for further research to enhance performance in such scenarios. The authors recommend exploring various unsupervised pretraining models and architectures for potential improvements and advocate for investigating the explainability and interpretability of the models' decisions. In summary, this paper showcases progress in identifying audio using unsupervised pretraining models, emphasizing their potential in constructing efficient systems for detecting deepfakes.

In [13], the paper introduces an audio anti-spoofing system designed to combat deepfakes and spoofing attacks, leveraging low-frequency sub-band information for robust detection. The system demonstrates effectiveness against a dataset containing deepfake audio examples, maintaining accuracy and resilience even in challenging scenarios with background noise and channel mismatch. While the paper presents a novel approach emphasizing low-frequency features, it acknowledges limitations, including a relatively small dataset, suggesting the need for broader evaluations and addressing potential vulnerabilities to advanced spoofing techniques. The authors recommend further research to enhance robustness and investigate computational efficiency for real-time applications.

In [14], the paper introduces a novel deepfake audio detection approach using bi-level optimization to enhance robustness against adversarial attacks. The method involves two optimization stages, with Level 1 utilizing a deep neural network for authenticity prediction and Level 2 employing an adversarial perturbation function. This iterative process results in a more resilient detection system, outperforming traditional models on a dataset of genuine and manipulated audio samples. Despite its promising potential, the methodology faces challenges such as potential

computational expenses and vulnerability to complex adversarial attacks. Further research is recommended for mitigating these issues and exploring interpretability. Overall, the paper offers a valuable advancement in creating robust audio verification systems for detecting deepfakes.

In [15], the paper introduces Quick-SpoofNet, a deep learning model designed for audio deepfake detection in voice anti-spoofing systems. Quick-SpoofNet employs innovative techniques, including one-shot learning, metric learning, and spectral feature analysis, to discern differences between real and manipulated audio. Its strength lies in its impressive ability to generalize effectively using minimal training data, achieving high accuracy and generalizability across various deepfake generation techniques. While contributing significantly to voice security, future research should involve testing on diverse real-world audio recordings, exploring different feature extraction methods, and assessing integration feasibility into existing voice biometric systems.

In [16], the paper presents SpecRNet, an innovative deep learning architecture for efficient audio deepfake detection. SpecRNet uses lightweight convolutional layers and residual blocks to reduce computational demands while maintaining high accuracy. It significantly reduces processing time, making it suitable for real-time applications, and demonstrates effectiveness across diverse datasets and conditions. Noteworthy strengths include faster and more accessible deepfake detection, especially in real-time scenarios, and compatibility with various devices. Future work should involve evaluation against real-world deepfakes, exploration of methods to enhance accuracy, and integration with existing audio processing pipelines and security systems. "SpecRNet" advances audio deepfake detection by offering a fast, efficient, and accurate model for practical use in safeguarding online communication.

In [17], the paper introduces the SE-Res2Net-Conformer architecture, a novel model designed for detecting synthetic voices and audio splicing. Combining SE-Res2Net for local pattern capture and Conformer for global temporal context, the model outperforms previous approaches in synthetic voice detection on the ASVspoof 2019 dataset. The paper also proposes a new formulation for audio splicing detection, emphasizing splicing segment boundaries, improving detection accuracy. While showcasing strengths in feature extraction and detection performance, the study acknowledges limitations in dataset size and suggests testing on more diverse datasets and real-world scenarios. The paper presents a promising approach to audio manipulation detection, offering improved performance in synthetic voice and spliced audio detection.

In [18], the authors propose a novel method for identifying deepfake audio using Mel-Frequency Cepstral Coefficients (MFCCs) and deep learning techniques. The approach leverages a deep neural network architecture, with CNNs outperforming other models in detecting manipulated audio on a controlled dataset. Despite considerable accuracy, the study recognizes limitations due to a smaller dataset, suggesting further research on larger and more diverse datasets to assess generalizability. The authors also recommend exploring different pre-processing techniques and evaluating the model's resilience against advanced deepfake generation methods and adversarial attacks. The paper marks a promising beginning in deepfake detection, acknowledging its limitations and paving the way for future advancements in more robust detection systems.

In [19], the paper introduces a novel deep learning approach for detecting fake audio messages using a hybrid model combining recurrent and convolutional neural networks (RNN-CNNs). The RNNs capture temporal dependencies, while CNNs extract spatial features from spectrograms. The model shows promising results on a dataset with real and fake audio messages, achieving high accuracy. The combined RNN-CNN approach outperforms using only RNNs or CNNs, highlighting its effectiveness in feature extraction and classification. While demonstrating innovative potential, the paper acknowledges limitations, including a relatively small dataset, suggesting further evaluation on larger datasets and exploration of different pre-processing techniques. Investigating the model's robustness against advanced deepfake techniques and adversarial attacks is suggested for future work.

In [20], the paper explores deep learning methods for detecting deepfake audio in digital forensics. It reviews existing deepfake audio classification methods and conducts a comparative analysis of various deep learning techniques, including custom architectures and pre-trained models like VGG-16. The evaluation considers features such as MFCC, Mel-spectrum, Chromagram, and spectrograms. Custom architectures excel with Chromagram, Spectrogram, and Mel-Spectrum features, while VGG-16 performs well with MFCC features. The paper contributes to forensic investigators' capabilities in distinguishing real and synthetic voices, offering insights for advancing digital forensics tools. Strengths include a comprehensive overview, method evaluation, and visualization of audio features. Limitations involve a limited dataset, suggesting further assessment with larger datasets, exploration of other feature sets, and investigation of model robustness against advanced deepfake techniques and adversarial attacks.

In [21], the paper introduces a novel deepfake detection approach by simultaneously analyzing audio and video modalities. This addresses the vulnerability of single-modality methods to manipulations targeting the other modality. The proposed deep learning architecture enables cross-modal interaction and information fusion, enhancing the model's ability to detect inconsistencies between audio

and video features. Evaluation on diverse deepfake datasets demonstrates effectiveness, with strengths including a multimodal approach and robust performance. However, the paper acknowledges dataset limitations, emphasizing the need for larger and more diverse datasets for further evaluation. It also highlights the importance of improving explainability and understanding the model's decision-making process, with further research needed for real-world scenarios involving sophisticated deepfakes and adversarial attacks.

In [22], the paper critically examines the impact of deepfakes on scientific knowledge dissemination and proposes mitigation strategies. The study emphasizes the susceptibility of individuals in the education sector to deepfake manipulation and underscores the need for detection tools, critical thinking skills, and information verification. A field experiment assesses vulnerability, revealing the necessity for targeted interventions and educational training. The study contributes by exploring this under-researched area and advocating a multi-pronged approach. Despite its strengths, the study has limitations, including a small sample size, necessitating further research with larger and more diverse populations. Future exploration of detection and mitigation strategies in real-world situations and the use of emerging technologies is suggested. Overall, the paper provides valuable insights into addressing deepfake impact on scientific information dissemination, advocating proactive measures for knowledge safeguarding.

3. Proposed Architecture

Fig. 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.

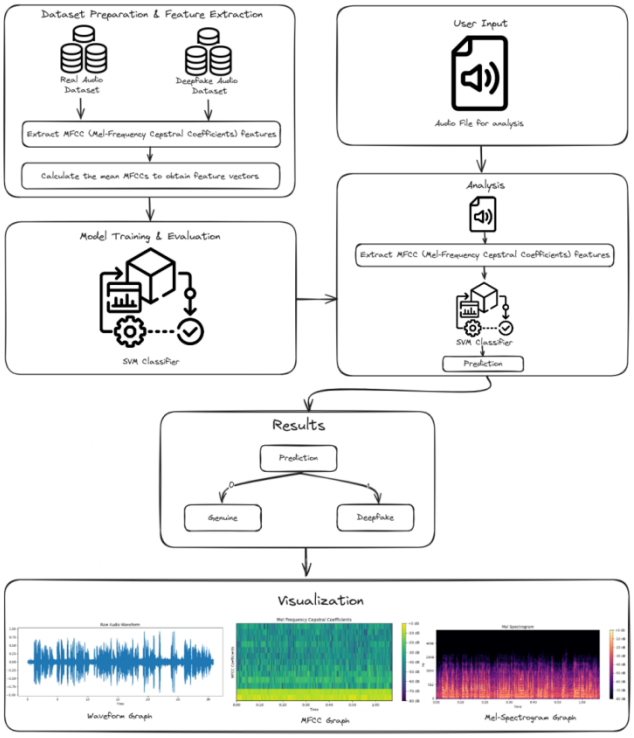

Fig. 1. Architecture of the Proposed Model

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.

Process of Computing MFCCs is explained in fig. 2. The process begins with the raw audio waveform, 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:

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

where is the Fourier Transform of the signal at frequency bin k, $x(n)$, is the $\boldsymbol{n-th}$ sample of the input signal, 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:

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

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

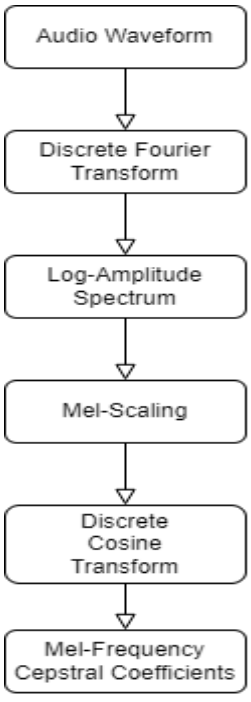

Fig. 2. Process of extraction of MFCCs

uniformly on the Mel scale which is shown in fig. 3. The filter bank output, $S(m)$, is given by:

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

where $H_m(k)$ is the Mel filter bank's $\boldsymbol{m-th}$ filter.

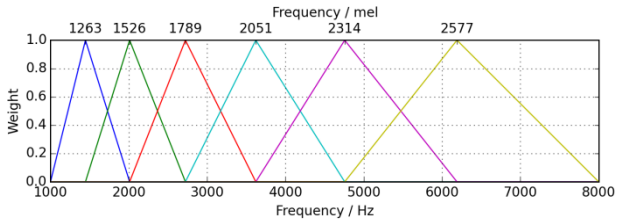

Fig. 3. Mel Scale

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:

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

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 $\boldsymbol{n-th}$ MFCC, $\boldsymbol{Cn}$, is calculated as follows:

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

for $\boldsymbol{n = 1, 2, \ldots, L}$, where $\boldsymbol{L}$ is the number of MFCCs kept for the analysis (often $\boldsymbol{L}$ 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.

The performance metrics extracted described in fig. 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

```
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]]
```

Fig. 4. Model Evaluation result

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.

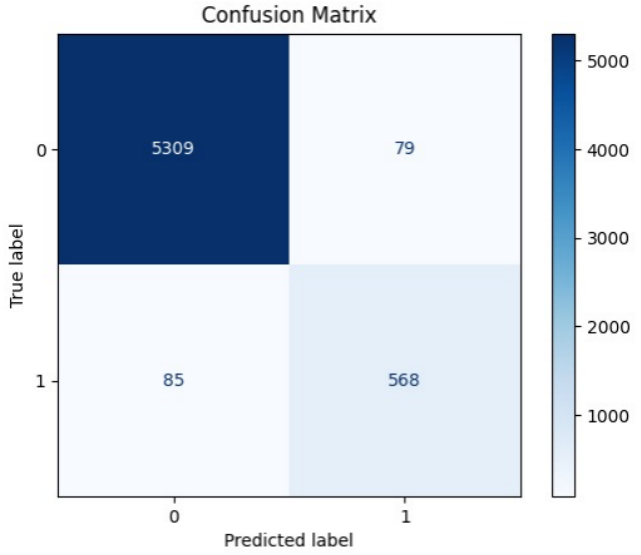

Fig. 5. Confusion Matrix for the model

Fig. 5 displays the confusion matrix for the test set.

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.

```
Enter the path of the Audio file: toanalyze\krishna.wav
The input audio is classified as genuine.

Enter the path of the Audio file: toanalyze\Ai Cloned.wav
The input audio is classified as deepfake.
```

Fig. 6. Result of input audio

The result shown in fig. 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

5.2.1. 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.

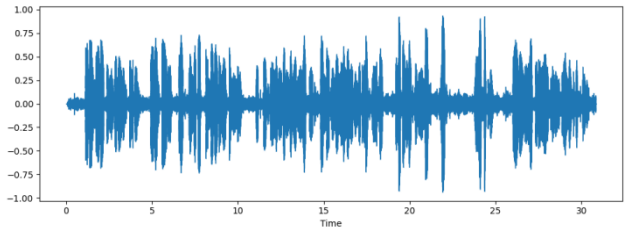

Fig. 7. Raw Real Audio Waveform

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

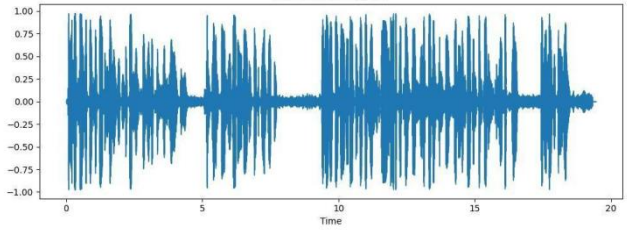

Fig. 8. Deepfake Audio Waveform

Fig. 8 illustrates the waveform of a deepfake audio sample.

5.2.2. 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.

In Fig. 9 & 10, we present two Mel spectrogram images: the first

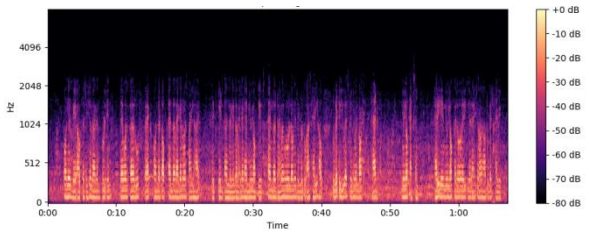

Fig. 9. Mel Spectogram of Real Audio

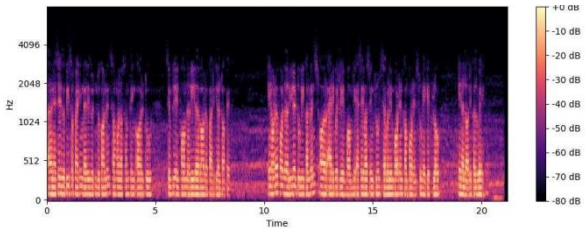

Fig. 10. Mel Spectogram of Deepfake Audio

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.

5.2.3. 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.

The comparison between the MFCC plots of the authentic and deepfake audio reveals distinct differences which are shown in fig. 11 & fig. 12.

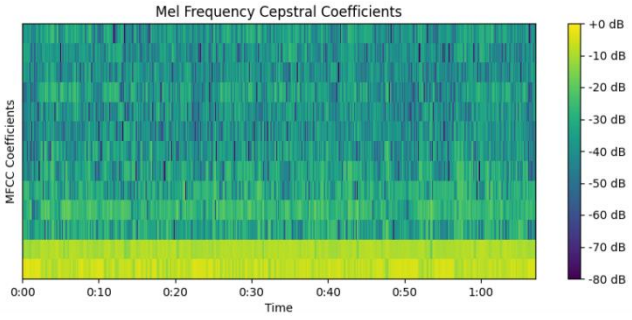

Fig. 11. MFCC graph for Real Audio

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

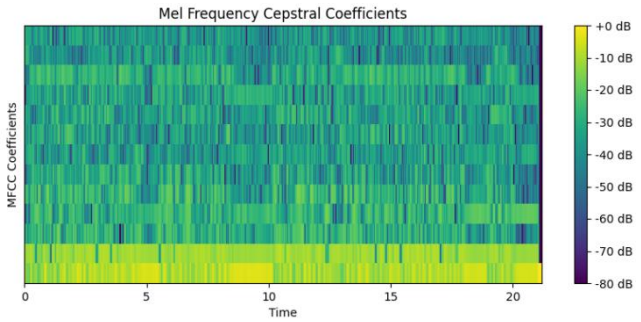

Fig. 12. MFCC graph for Deepfake Audio

plot displays irregularities and inconsistencies in the cepstral coefficients, indicating potential manipulation.

6. Conclusion

This research advances synthetic audio detection using a Support Vector Machine (SVM) classifier trained on the diverse 'for-original' Fake-or-Real (FoR) dataset. Leveraging Mel-Frequency Cepstral Coefficients (MFCCs) as features, the model achieves an impressive 97.28% accuracy, showcasing its robustness for digital authentication and security. The strategic use of the FoR dataset enhances performance in test environments and real-life scenarios, contributing significantly to cybersecurity and digital forensics. As deepfake threats evolve, tools like this play a crucial role in combating digital fraud and misinformation. Future work involves expanding the dataset, exploring advanced algorithms, and enhancing interpretability and user interface for broader applicability. In conclusion, this research provides a pivotal step in effective deepfake detection, laying the groundwork for ongoing efforts to preserve digital information integrity.

Acknowledgements

We express sincere gratitude to the Cyber Peace Foundation for their generous and unwavering support of our research project. Their dedication to propelling knowledge and innovation in the field has played a pivotal role in the successful realization of our work. Through their substantial financial backing, we have had the opportunity to further explore our research objectives, expanding the horizons of comprehension and making valuable contributions to both academic and practical discussions. This partnership reflects their steadfast commitment to promoting excellence in research, and we deeply appreciate the substantial impact of their contribution to our endeavors.

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