



Segment-Level Detection of Synthetic Voice Manipulation System

¹N. Manasa, ²Sk. Anisah Firdaws, ³P. Jeevan Sai, ⁴P. Kamal, ⁵P. Ravi Prakash

¹⁻⁴IV B. Tech, Department of Information Technology, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, India.

⁵Assistant Professor, Department of IT, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, India

Abstract: The synthetic voice generation technology has reached new heights because of artificial intelligence which has developed at an exceptional pace. The new technology for creating voices helps virtual assistants and accessibility tools work better. It also brings big security risks from two main sources: fake voice recordings and deepfake audio attacks. Scammers can now make realistic voice recordings that sound like real people. This makes it harder to tell if audio content is real or fake.

The threats are especially bad for Europe in areas like transactions, identity verification and social engineering. Detection systems currently look at the audio recording to decide if it's real or fake. They work well in some situations. Can't catch fake parts when only some of the audio is changed. This makes traditional systems less reliable in forensic investigations

The reason is that they often don't have real audio to work with. This paper introduces a solution: the Segment- Level Synthetic Voice Manipulation Detection system. It looks at signals in short intervals. The system breaks down the audio into parts and checks each one separately using deep learning. It uses Mel-Frequency Cepstral Coefficients and other features, which are then processed by a Convolutional Neural Network model. This produces probability results that show the chance of manipulation.

Index Terms - Segment-Level Detection, Synthetic Voice Detection, Deep Learning, Convolutional Neural Network, Mel-Frequency Cepstral Coefficients, Audio Forensics, Deepfake Audio, Signal Processing, Fraud Detection, Security Monitoring

I. INTRODUCTION

The proliferation of Artificial Intelligence (AI) and Deep Learning has transformed speech synthesis technologies, allowing for the development of extremely convincing fake/replica voices. While there are many useful applications of this technology, there are also many potential dangers associated with it, especially when it is used inappropriately. In particular, the manipulation of synthetic voices or audio deepfakes has become an important issue for cyber security and electronic communications.

The use of Artificial Intelligence and Deep Learning has changed speech synthesis technologies. They now allow for the creation of realistic fake voices. This technology has useful applications but it also has potential dangers. These dangers are especially present when it is used in the way. The manipulation of voices or audio deepfakes is a major issue. It affects security and electronic communications.

Current detection methods analyze a digital recording. They then give a simple yes or no answer. In real life most fake audio is only partly manipulated. This makes current detection systems not very effective.

For instance in a phone call only the parts with important financial information may be changed. The rest of the call is genuine. To fix the problems with methods we propose a new system called Segment-Level Synthetic Voice Manipulation Detection (SVMD).

SVMD uses an analysis approach. It breaks a recording into smaller parts and analyzes each part separately. This change will make detection more accurate. It will also give specific information about the fake parts of the audio. SVMD can also make it easier to understand the results of the audio analysis. It provides a report of all parts analyzed. This makes SVMD perfect for use in detecting fraud and analyzing evidence. The Artificial Intelligence and Deep Learning technology is very powerful. It can be used for good or bad. The Segment-Level Synthetic Voice Manipulation Detection system is a tool. It can help keep us safe from the uses of this technology. The SVMD system is designed to detect voices. It can help us know if a voice is fake or not. The use of SVMD can help prevent fraud. It can also help in evidence analysis.

II. NEED OF THE STUDY

As speech synthesis techniques have advanced rapidly, making it easier to make high-quality fake audio, they have been applied in various applications, including virtual assistants and content creation. Nevertheless, the same technology can be utilized to fabricate misleading fake audio clips that are very hard to differentiate from human voice recordings. The main problem with current fake audio detection solutions is that they deliver binary decision results concerning the authenticity of whole audio files. That means there is a need for a more advanced system to perform audio analysis with segment-level resolution.

Moreover, the increasing popularity of voice communication in banking operations and online interactions makes audio clips especially important since they can be used by fraudsters to impersonate their victims. It calls for sophisticated algorithms that are capable of differentiating between original and fabricated recordings. At the same time, it is important that such an algorithm provides information about what exactly was fabricated in audio files. Segment-level detection will improve its reliability and interpretability and will allow for expanding the range of possible application areas, including digital forensics.

III. RELATED WORKS

Current advancements in speech synthesis technology involve creating natural-sounding synthetic voices through the use of Artificial Intelligence and Deep Learning.

Current approaches to detecting synthetic speech typically involve examining an entire audio file and providing one output. While these approaches can be useful for certain tasks, they cannot handle situations where the manipulation is performed on only a portion of the audio file.

Audio classification algorithms like Mel-Frequency Cepstral Coefficients and Convolutional Neural Networks have been proven to be very effective. However, most models do not perform analysis at the

segment level, making them inefficient in detecting partially manipulated audio.

3.1 Literature Review

With the fast pace of development in speech synthesis and deep learning, the problem of detecting synthetic voice manipulation has gained more traction. Most researchers use either machine-learning models or signal- processing methods to detect deepfake audio. The current state-of-the-art methods seek to improve detection accuracy, robustness, and generalization performance across different datasets.

The contributions by Brendan McMahan et al. [1] have led to the emergence of decentralized learning, which has indirectly impacted the development of modern distributed artificial intelligence (AI) algorithms. Although the focus of McMahan's paper was to provide an overview of decentralized learning through examples of federated learning, local process execution can effectively address the privacy and efficiency problems of audio analysis systems.

Bonettini et al. [2] came up with the innovative approach of detecting deepfake audio using a convolutional neural network (CNN) algorithm based on spectrograms. In their research, they constructed a model that converts audio signals to spectrogram images, then use CNN architectures to detect whether the images represent real or fake audio signals. Their contribution holds tremendous potential for detecting deepfake audio.

The ASVspoof challenge was introduced by Yong Wang et al. [3] to develop benchmark datasets that will serve as a benchmark for testing detection systems designed to detect spoofed or synthesized voices. The contribution of these authors has been used as a benchmark for evaluating machine learning algorithms for training and testing models aimed at detecting deepfake audio. Deep learning models for detecting deepfake audio have been designed based on spectrogram features and ensemble learning algorithms by Long Pham et al. [4]. In this study, the researchers emphasized the significance of feature extraction methods like Mel Frequency Cepstral Coefficients (MFCC) and Spectral features in enhancing detection performance. The potential of transformers for end-to-end detection of deepfake audio voices has been studied by Lingyu Gong et al. [5]. The researchers found that using acoustic feature fusion along with transformer models enhances the performance and robustness of classifiers. A comparative study of deep learning models for detecting synthetic audio has been performed by Ali Verma et al. [6]. The authors suggested that CNN architectures offer an optimal solution to achieve an accurate and efficient model. Yue Xie et al. [7] have proposed a ResNeXt-based architecture using spectral features for detecting manipulated audio.

Their study highlights challenges such as generalization, dataset limitations, and robustness against new manipulation methods.

Despite these advancements, most existing systems focus on whole-audio classification and lack the ability to identify partially manipulated segments within an audio recording. This limitation reduces their effectiveness in real-world forensic applications.

Therefore, the proposed Segment-Level Synthetic Voice Manipulation Detection (SVMD) system addresses this gap by performing fine-grained temporal analysis, enabling detection of localized synthetic insertions and improving interpretability through segment-wise probability evaluation.

Table.1. Comparison Table

Feature	Previous Methods	Proposed System
Detection Approach	Whole audio analysis	Segment-level analysis
Output	Single real/fake result	Probability for each segment
Feature Extraction	Basic audio features	MFCC and advanced features
Model Used	Traditional ML models	CNN-based deep learning model
Manipulation Detection	Cannot detect partial changes	Detects partially manipulated segments
Interpretability	Limited	Detailed segment-wise results
Application Scope	General detection	Fraud detection, digital forensics, security monitoring

3.2 Comparison with Previous Methodology

Conventional audio manipulation detection systems mostly depend on the entire audio analysis process and generate one classification output for an audio file. Conventional audio manipulation detection systems overlook the fact that real-life audio files might have parts that were modified and not necessarily all. Consequently, these detection systems cannot detect manipulated audio segments, thus making them less efficient when applied to fields like fraud detection and digital forensics. In addition, current detection systems lack interpretability since they do not show the exact portion of the audio file that was tampered with.

Unlike the conventional systems, the proposed audio manipulation detection system based on deep learning (Segment-Level Synthetic Voice Manipulation Detection [SVMD]) applies a segment-based approach to produce more precise and informative outputs. The proposed system first splits the audio file into several segments before analyzing each segment individually. By employing Mel Frequency Cepstral Coefficients and Convolutional Neural Networks (CNN), the proposed system produces probability scores for each audio segment, indicating whether it is manipulated. Compared to other approaches, the proposed system is better at detecting manipulated segments, enhancing interpretability and providing practical applications in fraud detection, digital forensics, and security surveillance.

3.3 Proposed framework

The architecture of the SVMD system will be able to analyze and detect any synthetic voice manipulation in a granular manner. The workflow of the system starts by receiving an input audio file, which is preprocessed and segmented. Segmentation is done by dividing the input audio into small time-based segments. Feature extraction of these segments will then be performed by using different algorithms like Mel-Frequency Cepstral Coefficients.

These features will then be used as inputs for the Convolutional Neural Network algorithm to classify each segmented part as a real or synthetic voice. Probability scores will then be generated for each segmented voice, and they will be used to form a final report. In addition, visualization parts will be included in the architecture, such as signal analysis graphs and a timeline of segment authenticity.

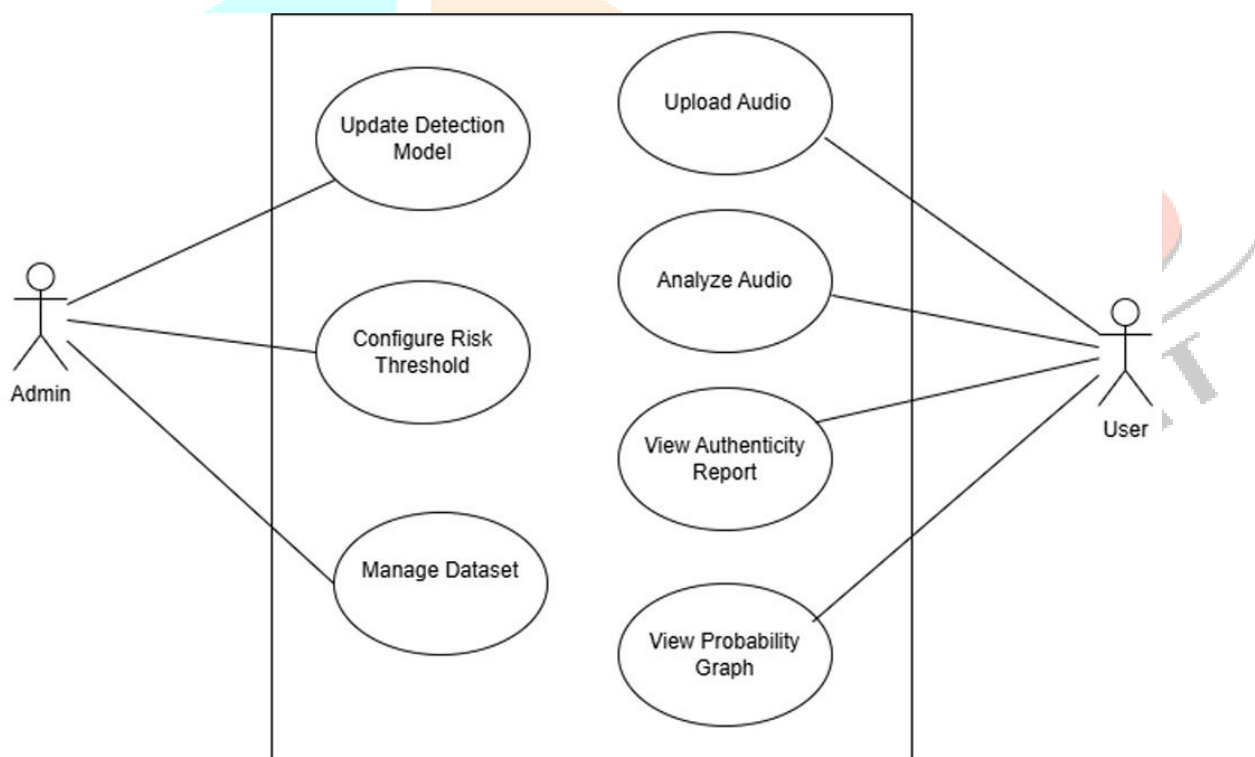
3.4 Main Methodology

The dedicated approach for detecting Segment-Level Synthetic Voice Manipulation (SVMD) involves examining entire segments of audio rather than whole pieces. Each segment is then analyzed separately so that we can determine whether one or more portions of the audio message are synthetically manipulated. The methodology consists of five steps as outlined below.

A. System Design & Structure

Initially, the system structure was developed around the architecture required to process audio input via multiple modules: an audio input interface, a preprocessing module, a segmentation module, a feature extraction module, a Deep Learning model, and a reporting module.

The user uploads an audio file using a web-based platform. After submission to the SVMD, the audio file is processed through validation and preprocessing steps before being segmented into fixed-length segments for sending as input into a trained Convolutional Neural Network (CNN) for classifying the audio segment as 'valid' or 'invalid.' The resulting classifications are then consolidated into a final authenticity.



report of the audio's

Fig. 1. Proposed SVMD system architecture.

B. Input Audio & Validating

Audio files accepted by SVMD are in the standardized audio file formats of either .wav or .mp3. After being uploaded by the user through the web platform, the system validates the format, duration, and quality of the audio file. If the audio file has an invalid file type or has been corrupted in some way, the system returns an error message to the user, asking them to upload an appropriate audio file.

C. Preprocessing Input Audio

After the audio has been validated, the audio undergoes many different types of preprocessing for

consistency purposes, which include:
Noise reduction (NR), Normalization (N),

Resampling to a standard sampling rate

Through the data preprocessing efforts, the quality of the audio signal has improved significantly and additional features extracted from the data become reliable in their representation.

D. Audio Segmentation

The audio is split into fixed-length pieces with a duration of two seconds per segment (the duration of some segments may vary) for evaluation, thus allowing the system to evaluate all the audio segments at once by independently evaluating each audio segment.

E. Feature Extraction

Acoustic characteristics of a segment will be identified for each segment, including:

- 1) Mel-frequency cepstral coefficients
- 2) Spectral features
- 3) Pitch-related features
- 4) Energy-related features

These features then serve as inputs to the deep learning model for classification.

F. Segment-Level Classification

All audio segments are then passed through a CNN Classifier (Convolutional Neural Network) for classification. The CNN classifies the extracted features of the audio segments and assigns it a probability score indicating whether or not it is a real audio segment or a synthetic audio segment.

G. Probability Aggregation

During this stage, the audio segments will generate a probability score between zero and one. Once all segments have been evaluated, the system combines the probability scores for each of the segments to create an overall authenticity score for the entire audio file.

Overall authenticity score is calculated as follows:

$$\text{Overall authentication score} = (P_1 + P_2 + \dots + P_n) / n$$

where P_n is the probability of each audio segment and n is the total number of audio segments processed.

H. Classifying Risk

According to the combined probability score, the audio is classified by the algorithm as follows: Low

Risk- Mostly Authentic

Medium Risk - Only Partially Altered High Risk - Largely Altered

The probability scores associated with the levels above are determined by preset thresholds which can be modified according to the technical requirements of the algorithm.

I. Reporting & Data Visualization

The report generated by the algorithm is a complete letter authenticating the audio with: Overall combined probability of audio being authentic

The likelihood of each form of manipulation occurring Risk level classification for each type of manipulation Probability for each manipulation compared to time

Disabled areas of content identified as possibly being altered.

The report and information generated above are clearly represented in an interpretable way, making the algorithm useful for providing comprehensive forensic analysis of audio files.

J. Benefits of the Current Methodology

There are numerous benefits associated with the current methodology including:

- 1.) the ability to identify partially altered audio recordings
- 2.) the ability to provide a segmented analysis of sound recordings, thus increasing the see-through abilities of investigators
- 3.) increased accuracy for sound manipulation compared to using complete and indivisible audio recordings
- 4.) possess the capability to generate comprehensive forensic-style reporting, enabling investigators to make objective decisions
- 5.) the capacity to perform audio manipulation analysis in real-time and/or large-scale audio manipulation cases.

IV. RESULTS AND DISCUSSION

4.1 System output screenshots and explanation

This section will show the results obtained from the Deep Signal Profiling system for Segment-Level Synthetic Voice Manipulation Detection. The primary goal was to test the ability of the system to provide accurate predictions at a segment level instead of providing only one general prediction output.

A. Experiment Configuration

The experiment involved the use of Deep Signal Profiling software interface for uploading audio files (.wav, .flac, .mp3). The uploaded audio input will be processed and segmented into smaller time intervals. Afterward, the individual segments were analyzed using their extracted features. They were also subjected to evaluation through a Convolutional Neural Network model. The system generated probability scores for each segment that determine its classification.

B. Global Model Prediction

Upon completing the analysis of the uploaded audio, the system provided a final prediction result. Based on the output, the model achieved high accuracy in predicting manipulated audio segments. In this specific output, the model identified the input audio as synthetic with a probability of 97.63%.

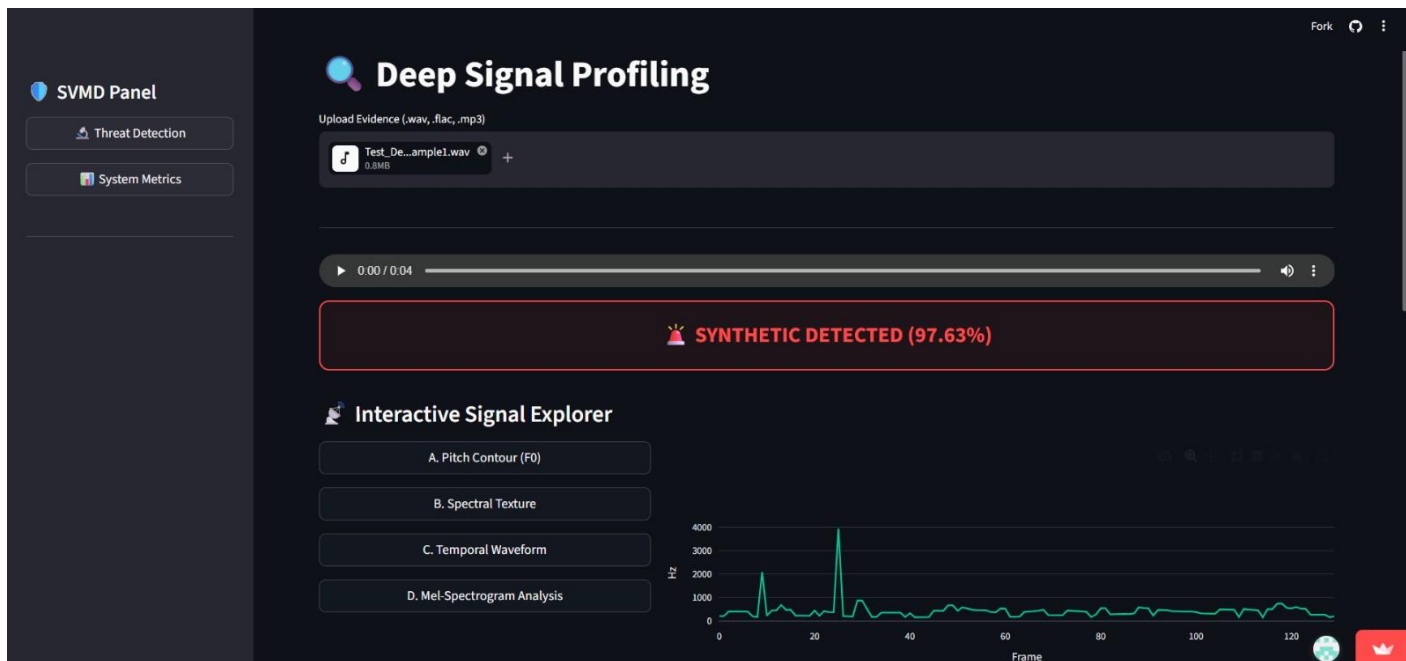


Fig. 2. Deep Signal Profiling Interface with Synthetic Detection Outcome

C. Signal Analysis and Feature Discovery

There is also an interactive signal explorer in the system that can be used to analyze various properties of the input including pitch contour, spectrogram pattern, temporal pattern of the sound wave, and Mel-spectrogram analysis. As per the results of the analysis, natural voice samples contain chaotic micro-fluctuations, whereas smooth pitch contour is an indication of artificial synthesis.



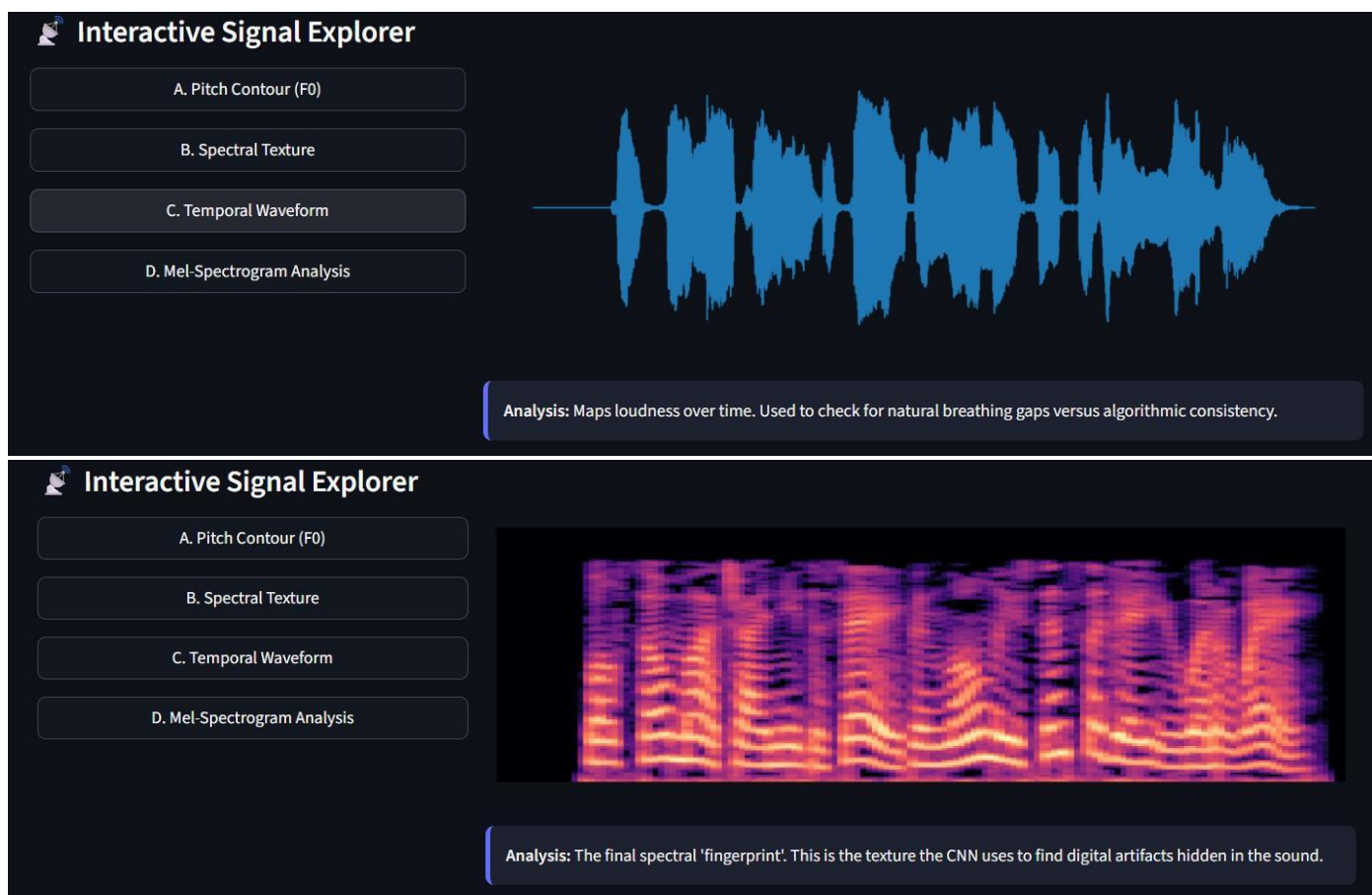


Figure 3. Interactive Signal Explorer with Pitch Contour Analysis

D. Segment Analysis

For detailed analysis, the audio is broken down into smaller segments and analyzed separately. From the authenticity graph at the segment level, it becomes clear that all segments of the input audio have very high values for fake probability. Hence, we can conclude that the manipulation of the audio is done uniformly throughout the audio file.

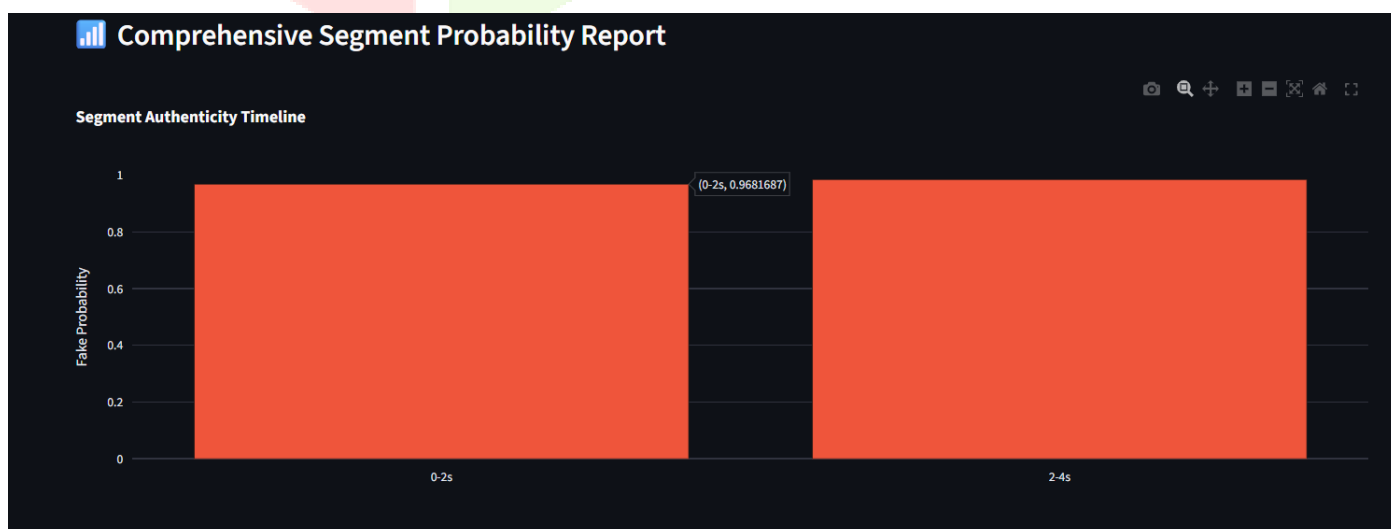


Figure 4. Segment Authenticity Graph

E. Detailed Segment Results

In addition to the above results, there is a section called 'Detailed Segment Results'. Here, time ranges, probability scores, and conclusions are presented along with other information. From the details shown, it was seen that the two analyzed segments in the present case were fake, with probability scores 96.82% and 98.44% respectively.

Detailed Segment Data			
Segment Index	Time Range	Fake Probability	Verdict
Seg 1	0-2s	96.82%	⚠️ FAKE
Seg 2	2-4s	98.44%	⚠️ FAKE

Figure 5. Detailed Segment Information with Probability Scores

F. Metrics for Assessing the Overall Performance of the System

Performance of the overall system has been assessed on the basis of common metrics. The accuracy of the model was estimated at 94.2%, while the precision stood at 92.8% and recall at 95.1%. As shown by the confusion matrix, most of the predictions made were correct, although some of them could be considered as being false positive or false negative. This suggests that the algorithm developed works properly to discriminate real from synthetic sounds.

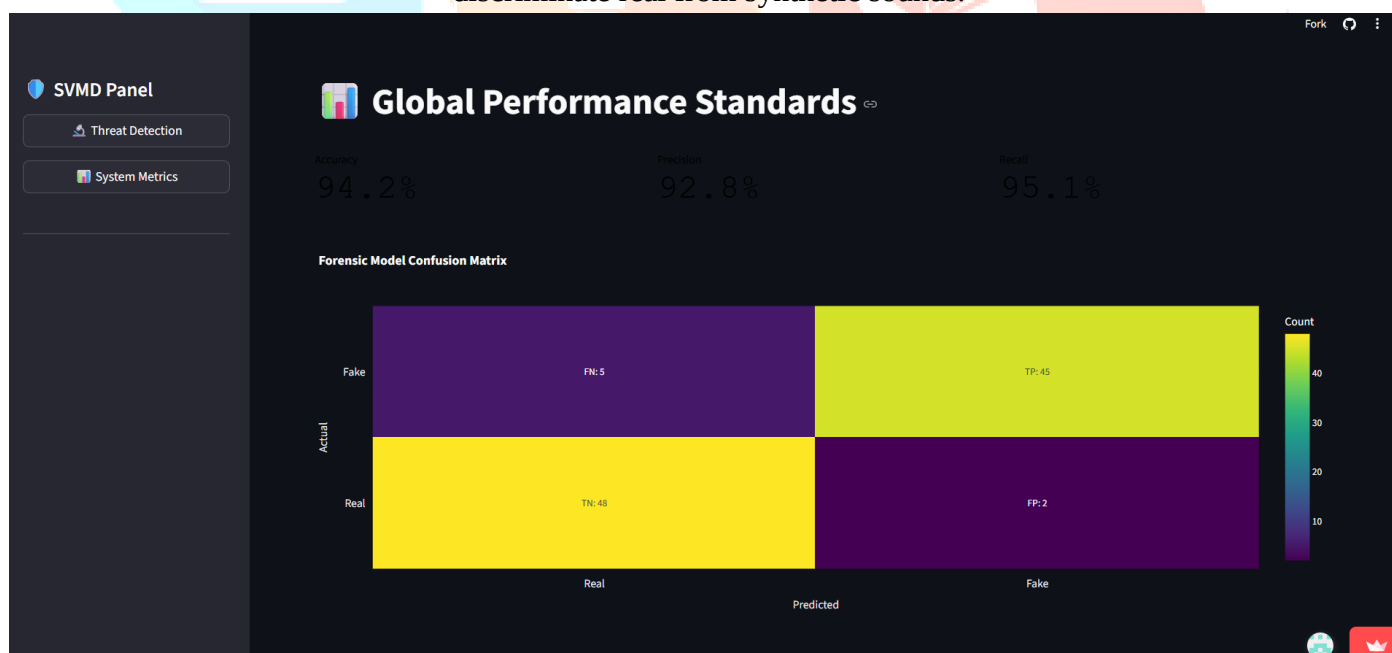


Fig. 6. Performance Metrics and Confusion Matrix

G. Predictions

Finally, the developed model was applied to perform prediction analysis based on previously unseen input audio samples. In the result, the system was able to provide reliable predictions together with segment-level analysis.

4.2 Conclusion

In this project, we developed a synthetic voice manipulation detection (SVMD) system that is able to detect audio recordings that were manipulated using a synthetic voice (i.e. a voice created by software). The primary goal of this system is not only to detect these manipulations at the overall audio level, but also to detect them at the finer granularity of segments of the audio (i.e. smaller pieces of audio). To accomplish this, the audio is divided into smaller segments and each segment is processed separately to overcome the limitations of whole-audio classification.

In addition to segmenting the audio, our proposed solution also extracts various acoustic features from the audio, including Mel-frequency cepstrum coefficients (MFCCs), spectral features, and pitch variations, and then uses a deep learning based model to classify each segment as either real or manipulated. The predictions from segmenting the audio are aggregated to create an overall judgement of the authenticity of the audio as well. Furthermore, detailed forensic reports are provided, which include segment-level probability and risk scores, as well as a timeline view of the audio, improving the interpretability of the results and allowing for better identification of suspicious regions of the audio.

The results of our experiments demonstrate that using a segment by segment analysis increases the accuracy of detecting manipulated audio recordings by providing better insights into the results than previously available methods. The system also successfully identified audio recordings that were only partly manipulated, which would not have been detected by the existing methods. Overall the proposed SVMD system enhances the reliability, transparency and usability in application areas such as fraud detection, digital forensics and security monitoring.

4.3 Future Scope

- Integration with Advanced Models – Expand the system by incorporating more advanced deep learning models to improve detection performance.
- Real-Time Audio Analysis – Implement real-time processing to analyze live audio streams and detect synthetic manipulation instantly.
- Improved Feature Extraction – Use additional signal processing techniques along with Mel-Frequency Cepstral Coefficients for better accuracy.
- Multilingual Audio Support – Extend the system to handle multiple languages and different speech patterns.
- Enhanced Detection Accuracy – Integrate more robust models and larger datasets to improve classification and reduce errors.
- Mobile Application Development – Develop a mobile-based interface for easier access and real-time audio verification.
- Integration with Security Systems – Apply the system in authentication systems, call verification, and monitoring applications.
- Personalized Alerts and Reporting – Provide detailed reports and automated alerts for detected manipulated audio segments.

REFERENCES

- [1] A. Verma et al., “Deepfake Audio Detection: A Comparative Study of Advanced Deep Learning Models,” IEEE Access, 2025.
- [2] Y. Xie, X. Wang, Z. Wang, and L. Ye, “Deepfake Audio Detection with Spectral Features and

ResNeXt-Based Architecture,” Knowledge-Based Systems, 2025.

[3] L. Pham, D. Tran, F. Skopik, A. Schindler, et al., “DIN-CTS: Low-Complexity Depthwise-Inception Neural Network with Contrastive Training Strategy for Deepfake Speech Detection,” arXiv preprint, arXiv:2502.20225, 2025.

[4] L. Pham, P. Lam, T. Nguyen, H. Nguyen, and A. Schindler, “Deepfake Audio Detection Using Spectrogram-Based Feature and Ensemble of Deep Learning Models,” arXiv preprint, arXiv:2407.01777, 2024.

[5] M. Aleem, S. Riaz, M. T. Aziz, and A. R. Chishti, “AI-Based Deepfake Audio Detection – A Review,” Spectrum of Engineering Sciences, 2025.

[6] L. Y. Gong and X. J. Li, “Deepfake Voice Detection: An Approach Using End-to-End Transformer with Acoustic Feature Fusion,” Electronics, vol. 14, no. 10, 2025.

[7] Y. Wang et al., “ASVspoof: Automatic Speaker Verification Spoofing and Countermeasures Challenge,” IEEE Transactions on Information Forensics and Security, 2019.

[8] N. Bonettini et al., “Detecting Deepfake Audio with Spectrogram-Based CNN,” in IEEE International Workshop on Information Forensics and Security (WIFS), 2020.

