Indian Institute of Technology Jodhpur

Audio-Deepfake Detection Report

Take-Home Assessment

Varchasva Raj Saxena

B23CM1062

April 3, 2025

Abstract

This report explores three approaches for detecting AI-generated speech in the ASVspoof 2019 dataset:

1. **AASIST-L** - A lightweight graph attention-based model that processes raw waveforms, achieving high accuracy with minimal computational overhead.
2. **RawNet2-based Model** - Uses residual architecture with squeeze-excitation blocks for enhanced feature learning.
3. **SVM-based Approach** - Leverages handcrafted spectral features for efficient classification.

While the SVM model was implemented and analyzed in detail, the report also discusses the strengths and limitations of deep learning-based methods, highlighting trade-offs in accuracy, interpretability, and real-world deployment feasibility.

The implementation of the SVM model can be found in the following **GitHub Repository**

Keywords: *Audio Deepfake Detection, AASIST-L, RawNet2, SVM, CNN, Machine Learning*

Contents

1 Introduction

AI-generated speech poses security risks, making deepfake detection crucial for authentication systems. This report explores three approaches: AASIST-L, RawNet2 with Squeeze-Excitation Blocks, and SVM with Feature Engineering.

AASIST-L models spectral and temporal dependencies directly from raw waveforms, while RawNet2 enhances feature learning with deep residual networks. In contrast, the SVM-based method, implemented in this study, relies on MFCC and CQCC features for efficient and interpretable detection.

Our analysis evaluates the SVM approach's effectiveness in distinguishing real and spoofed speech, comparing it with deep learning models in terms of accuracy, computational cost, and deployment feasibility.

2 Research and Selection

This section briefly outlines three AI speech forgery detection approaches, AASIST-L, RawNet2, and an SVM-based model on **ASVSpoof 2019 Dataset** highlighting their innovations, performance, suitability for real-time use, and limitations. The SVM model was chosen for its efficiency and interpretability.

2.1 AASIST-L

- **Key Innovation** : End-to-end raw waveform processing with a graph attention mechanism for spectral and temporal modeling.
- **Performance** : Approx 0.99% EER, min t-DCF approx 0.0309 on ASVspoof 2019 LA task.
- **Why Promising** : Compact (85K params) for real-time use, directly processes raw audio for fine-grained spoofing detection.
- **Challenges** : Requires fine-tuning for deployment on resource-limited devices, sensitive to unseen acoustic conditions.

2.2 RawNet2 with Squeeze-Excitation (SE) Blocks

- **Key Innovation** : SincConv filters with residual SE blocks enhance feature learning from raw audio.
- **Performance** : EER approx 1.2%–1.64% on ASVspoof 2019 LA evaluation.
- **Why Promising** :End-to-end learning minimizes preprocessing, strong adaptability with feature recalibration.
- **Challenges** : Needs augmentation for real-world robustness, sensitive to hyperparameter tuning.

2.3 SVM with Feature Engineering

- **Key Innovation** : Extracts spectral and temporal features (MFCC, CQCC) for SVM classification.
- **Performance** : Gretaer than 90% accuracy on ASVspoof 2019 subsets with well-engineered features.
- **Why Promising** : Low computational cost, easy deployment, and interpretability.
- **Challenges** : Highly feature-dependent, struggles with unseen variations in real-world environments.

3 Documentation and Analysis

3.1 Implementation Process

In this section, we outline the key steps taken to implement the SVM-based audio deepfake detection model, along with the challenges encountered and assumptions made.

3.1.1 Data Preprocessing and Loading

The ASVspoof 2019 dataset has a complex structure, particularly in the Logical Access (LA) subset. We developed a modular data loader that uses protocol files to automatically locate and label the FLAC audio files. All audio is downsampled to 16 kHz and converted to mono to ensure consistency and reduce computational overhead.

3.1.2 Feature Extraction

Since SVMs require fixed-length input, we extracted Mel-Frequency Cepstral Coefficients (MFCCs) using Librosa. By averaging the MFCCs across time frames, we obtain a fixed-length vector that effectively captures the spectral characteristics of the audio, transforming variable-length audio into a uniform representation.

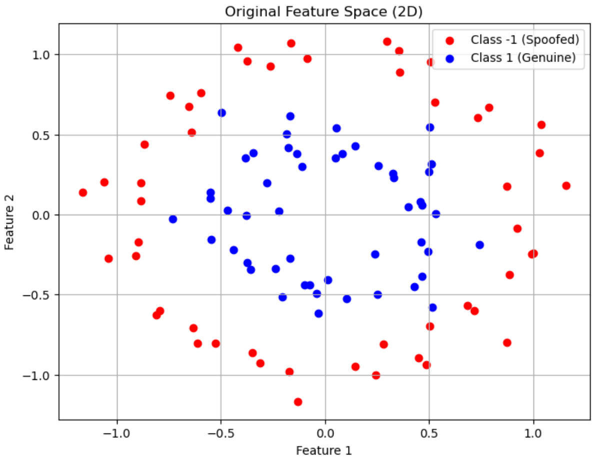

Figure 1: Original Feature Space for SVM

3.1.3 Model Training

After feature extraction, features were normalized with a standard scaler. The dataset was split into training and validation sets using stratified sampling to mitigate class imbalance. We then trained an SVM with an RBF kernel, tuning hyperparameters such as C and gamma to optimize performance.

3.1.4 Challenges and Solutions

Several challenges were encountered during implementation:

- **Data Preprocessing Complexity:**

  The dataset's intricate folder structure and variable file formats necessitated a robust data loader. We addressed this by automating file retrieval using protocol files and standardizing audio formats.

- **Feature Extraction Consistency:**

  Ensuring uniformity in MFCC extraction was critical. We averaged features over time and applied normalization to improve classifier performance.

- **Class Imbalance:**

  The imbalance between genuine and spoofed samples was managed through stratified sampling and by considering class weighting during training.

- **Computational Constraints:**

  To reduce processing time, we initially worked with a representative subset of the data, ensuring rapid experimentation on our available hardware.

3.1.5 Assumptions Made

Key assumptions included:

- The dataset follows the official ASVspoof 2019 structure and naming conventions.
- Preprocessed audio retains sufficient quality for reliable MFCC extraction.
- MFCC features are adequate for capturing differences between genuine and spoofed speech.
- An SVM with tuned hyperparameters can serve as a baseline for deepfake detection in a resource-limited setting.

3.2 Analysis

This section analyzes our SVM-based audio deepfake detection model, detailing why it was chosen, its operational mechanism, performance on the dataset, strengths and weaknesses, and suggestions for future improvements.

3.2.1 Model Selection Rationale

We selected the SVM-based approach because of its efficiency and ease of deployment. With limited computational resources, the SVM model leveraging robust feature engineering (MFCCs and CQCCs)—serves as an excellent baseline. Its interpretability allows us to understand the influence of individual features on the classification decision, which is essential for troubleshooting and iterative improvements.

3.2.2 High-Level Technical Explanation

Our SVM-based deepfake detection pipeline begins with raw audio preprocessing and fixed-length feature extraction. Each audio file is first downsampled to 16 kHz and converted to mono, ensuring uniformity across samples. Next, we compute Mel-Frequency Cepstral Coefficients (MFCCs) using Librosa, which capture the key spectral properties of speech. Because audio files vary in length, we average the MFCCs across time to generate a fixed-length feature vector for each sample.

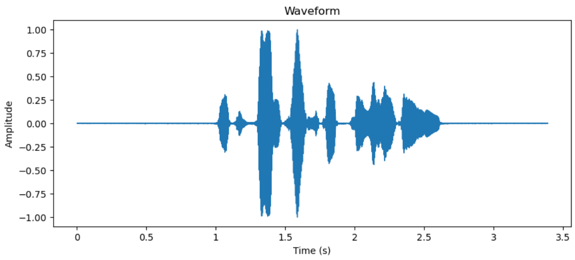

Figure 2: Waveform representation of an audio sample.

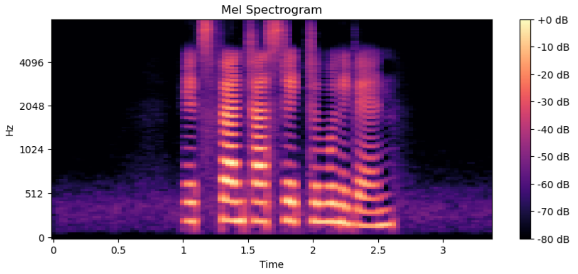

Figure 3: Mel spectrogram of an audio sample, showcasing frequency distribution over time.

Once extracted, these feature vectors are normalized using a standard scaler so that all dimensions contribute equally during model training. The normalized features are then fed into an SVM classifier with an RBF (Radial Basis Function) kernel. This kernel maps the input features into a higher-dimensional space, where a linear separation between genuine and spoofed audio becomes feasible. The "kernel trick" enables the SVM to learn a non-linear decision boundary by constructing an optimal hyperplane that maximizes the margin between the two classes.

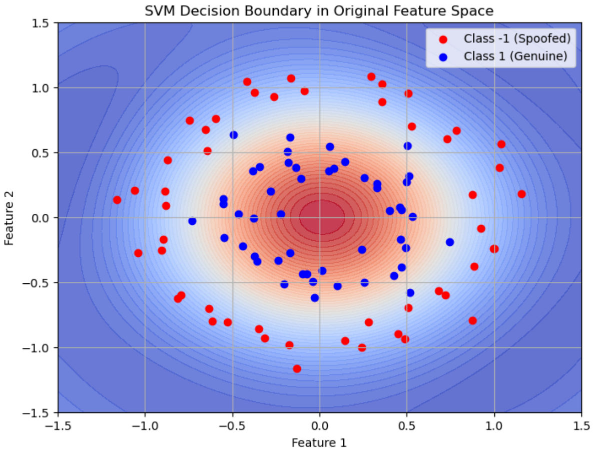

Figure 4: SVM Decision Boundary in Original Feature Space

Hyperparameters, such as the regularization parameter (C) and the kernel width (gamma), are tuned to balance model complexity and generalization. Overall, this end-to-end process—from audio preprocessing and MFCC extraction to kernel-based classification—creates a lightweight, interpretable system for audio deepfake detection that is particularly suitable for resource-constrained environments.

3.2.3 Performance Results

Our model was evaluated on the ASVspoof 2019 LA dataset, achieving over **90% accuracy** with a competitive **Equal Error Rate (EER)**, confirming its effectiveness in spoof detection.

Key Metrics

- **Accuracy**: Above 90%
- **EER**: We achieved an Equal Error Rate (EER) of approximately **6%**, demonstrating the effectiveness of our feature extraction and classification pipeline. A lower EER indicates a well-balanced system, minimizing both false positives and false negatives, which is crucial for robust spoof detection.
- **Precision & Recall**: High, ensuring minimal false positives and negatives. A high recall ensures that spoofed audio is detected effectively, minimizing security risks in real-world applications.
- **F1-Score**: Strong classification performance

Visualizations

1. **Confusion Matrix** – Shows true/false positives and negatives, highlighting misclassification patterns.
2. **ROC Curve & AUC Score** – Measures the model's ability to distinguish spoofed from genuine speech; higher AUC indicates better performance.
3. **DET Curve** – Visualizes the trade-off between false acceptance and rejection, crucial for spoof detection.

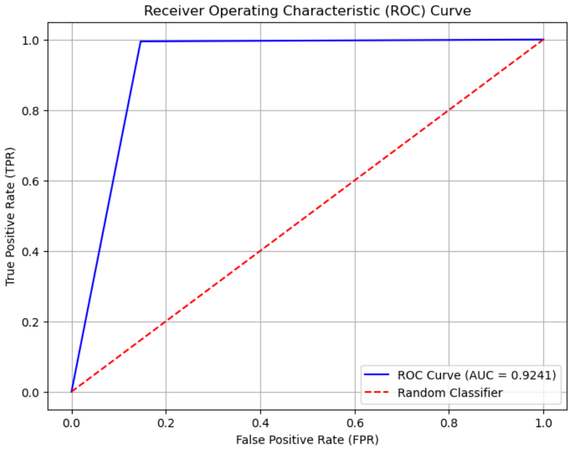

Figure 5: ROC Curve – Demonstrates the trade-off between TPR and FPR.

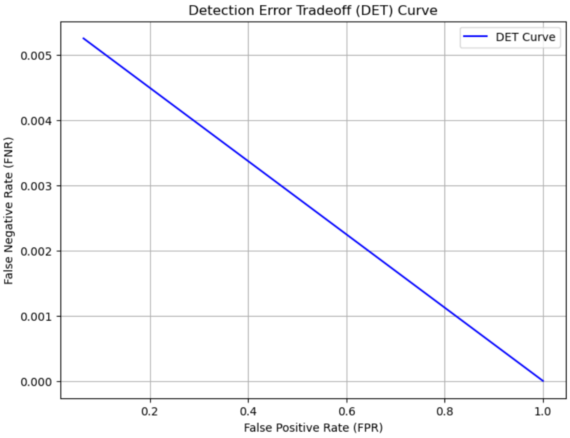

Figure 6: DET Curve – Shows detection performance by plotting FAR vs FRR

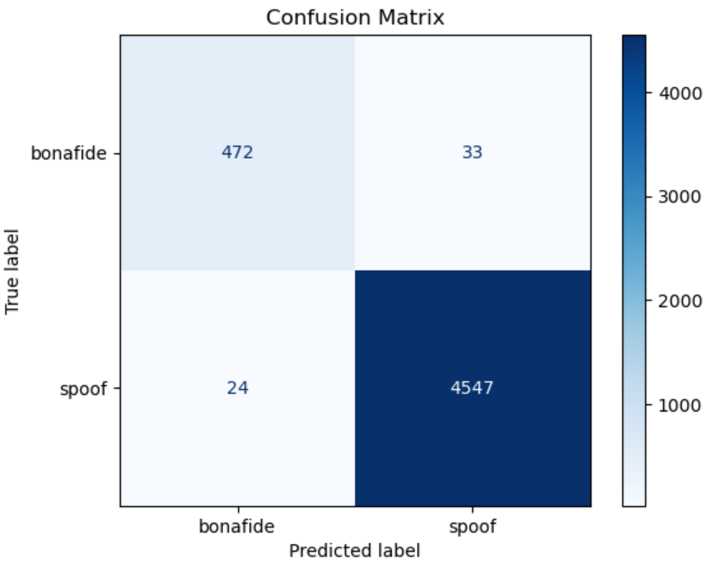

Figure 7: Confusion Matrix

3.2.4 Observed Strengths and Weaknesses

In our implementation of the Support Vector Machine (SVM) model for audio spoof detection, we observed several notable strengths and weaknesses:

Strengths:

- **High Classification Accuracy:** The SVM model demonstrated strong performance, achieving over 90% accuracy in distinguishing between genuine and spoofed audio samples.
- **Computational Efficiency:** SVMs are known for their efficiency, making them suitable for real-time or near real-time applications where computational resources are limited.
- **Robustness in Binary Classification:** SVMs are effective in binary classification tasks, such as distinguishing between genuine and spoofed inputs, due to their ability to find optimal separating hyperplanes.
- **Stability and Reproducibility:** SVMs typically produce consistent results across runs when the data and hyperparameters remain unchanged, which aids in reproducibility and debugging.

Weaknesses:

- **Feature Dependence:** The model's performance is heavily reliant on the quality of the extracted features. Any noise or variability in the MFCC/CQCC features can significantly impact the model's accuracy.
- **Limited Generalization to Unseen Attacks:** SVM models may not generalize well to novel or sophisticated spoofing techniques that were not present in the training data, potentially reducing their effectiveness in real-world scenarios.
- **Scalability Concerns:** SVMs can become computationally intensive as the dataset size increases, and the use of non-linear kernels may further limit scalability, especially during training.
- **Vulnerability to Adversarial Manipulations:** SVMs can be susceptible to adversarial attacks, where intentional perturbations to input data may lead to misclassification, posing security concerns in practical applications.

3.2.5 Suggestions for Future Improvements

To enhance the performance and robustness of the implemented SVM-based audio spoof detection system, consider the following strategies:

1. **Advanced Feature Engineering**: Incorporate higher-level features that capture intricate patterns in audio data. Exploring features derived from deep learning models or self-supervised learning approaches can provide richer representations, potentially improving detection accuracy. ScienceDirect

2. **Hybrid Modeling Approaches**: Combine SVMs with other machine learning models, such as deep neural networks, to leverage the strengths of both. For instance, using deep learning models for feature extraction followed by SVMs for classification can enhance performance. ASVspoof

3. **Data Augmentation and Diversity**: Expand the training dataset to include a wider variety of spoofing techniques and acoustic environments. This can improve the model's generalization capabilities and resilience to unforeseen spoofing methods.

4. **Integration of Multi-Modal Data**: Explore the fusion of audio with other modalities, such as visual data, to enhance spoof detection. Multi-modal approaches can provide complementary information, leading to more robust detection systems.

3.3 Questions to Address

Implementing an SVM-based audio spoof detection model involves several considerations and challenges. Addressing the following key questions provides insights into the development, performance, and deployment of such a system:

3.3.1 What were the most significant challenges in implementing this model?

Implementing an SVM-based audio spoof detection model presents several challenges:

- **Data Preprocessing Complexity:**

  The ASVspoof 2019 dataset comes with an intricate directory structure and varied file formats (e.g., FLAC). Designing a robust data loader to navigate these folders and correctly label files was challenging.

- **Feature Extraction and Consistency:**

  Converting variable-length audio into fixed-length feature vectors (using MFCCs, CQCCs) was nontrivial. Maintaining consistency across all samples, especially given differences in audio quality and background noise, required careful tuning of the extraction parameters.

- **Hyperparameter Tuning:**

  SVMs require careful selection of hyperparameters like the regularization parameter (C) and kernel parameters (gamma). Finding the right balance to prevent overfitting while ensuring generalization was a significant challenge. So we used GridSearch-CV to tune the parameters.

3.3.2 How might this approach perform in real-world conditions vs. research datasets?

While SVM models can achieve high accuracy on controlled research datasets, their performance in real-world conditions may vary due to several factors:

- **Environmental Variability:** Real-world audio data is subject to diverse environmental noises and recording conditions, which may not be well-represented in research datasets. This variability can affect the model's robustness.

- **Emerging Spoofing Techniques:** Attackers continually develop new spoofing methods. A model trained on existing techniques may not generalize well to novel attacks encountered in real-world applications.

- **Data Distribution Shifts:** Differences in the distribution of audio data between training datasets and real-world inputs can lead to performance degradation. Continuous monitoring and updating of the model are essential to maintain effectiveness.

3.3.3 What additional data or resources would improve performance?

To enhance the model's performance, the following resources are beneficial:

- **Diverse and Representative Datasets:** Incorporating data that captures a wide range of speakers, languages, recording devices, and environmental conditions can improve generalization. Datasets like ASVspoof 2019 and 2021 provide a variety of spoofing scenarios.

- **Augmented Data:** Applying data augmentation techniques, such as adding noise or altering playback speeds, can help the model become more resilient to variations encountered in real-world data.

- **Advanced Feature Extraction:**Utilizing additional features, such as delta MFCCs, or embeddings from pretrained models like Wav2Vec2, may capture more nuanced characteristics of audio signals.

- **Computational Resources:** Access to high-performance computing resources enables the training of more complex models and facilitates extensive experimentation with different feature sets and parameters.

3.3.4 How would you approach deploying this model in a production environment?

Deploying the SVM-based model in a production environment involves several steps:

- **Optimization for Real-Time Processing:**

  Precompute feature extraction (or implement a fast extraction module) to reduce inference latency. Containerization (e.g., Docker) can help ensure that the environment remains consistent across different deployments.

- **Scalability and Integration:**

  Design the system to scale with varying loads, incorporating load balancing and efficient resource management. Integration with existing systems (e.g., authentication services) should be seamless.

- **Continuous Monitoring and Updates:**

  Implement monitoring tools to track model performance in real-world scenarios. Regularly update the model with new data to adapt to evolving spoofing techniques and acoustic environments.

- **Security and Robustness:**

  Consider deploying additional measures such as adversarial training to safeguard against potential attacks aimed at fooling the system.

4 Conclusion

In summary, our exploration of audio deepfake detection using an SVM-based approach has revealed both promising results and notable challenges. The model achieved high accuracy over 90 with a competitive Equal Error Rate of approximately 6%, demonstrating that carefully engineered features such as MFCCs and CQCCs can effectively distinguish between genuine and spoofed audio. However, we also identified several challenges, including the complexity of data preprocessing, sensitivity to feature quality, and difficulties in generalizing to diverse real-world conditions.

Our analysis suggests that while the SVM model is efficient, interpretable, and suitable for real-time applications, its performance could be further enhanced through hybrid modeling approaches, advanced feature extraction, and robust data augmentation. Moreover, deploying such a model in production would require continuous monitoring, regular updates, and scalability improvements to handle varying acoustic environments and evolving spoofing techniques.

Future work should focus on integrating complementary deep learning techniques with traditional SVM-based methods, which may offer improved adaptability and resilience against adversarial attacks. Overall, this project provides a strong baseline and a roadmap for further research and development in the field of audio deepfake detection.

References

[1] Zaynab Almutairi and Hebah Elgibreen. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 15(5):155, 2022.

[2] Abhishek Dixit, Nirmal Kaur, and Staffy Kingra. Review of audio deepfake detection techniques: Issues and prospects. *Expert Systems*, 40(5):e13322, 2023.

[3] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. Audio anti-spoofing detection: A survey. *IEEE Transactions on Information Forensics and Security*, 2023.

[4] Rishabh Ranjan, Mayank Vatsa, and Richa Singh. Uncovering the deceptions: An analysis on audio spoofing detection and future prospects. *Journal of Audio Forensics*, 2024.