

# VoxSentinel : Voice-Based Authentication System with Spoofing Attack Detection

Vaibhavi Dalvi<sup>1</sup>, Ariba Shaikh<sup>2</sup>

<sup>1</sup>Vaibhavi Dalvi, Dept. of Data Science Engineering, Usha Mittal Institute of Technology, Mumbai, India

<sup>2</sup>Ariba Shaikh, Dept. of Data Science Engineering, Usha Mittal Institute of Technology, Mumbai, India

\*\*\*

**Abstract** - Automatic Speaker Verification (ASV) systems have grown widely adopted in sectors such as banking, smart devices, and secure access due to their ease of use and contactless operation. However, these systems remain vulnerable to various spoofing attacks, including replay attacks using recorded voices, synthetic speech generated via Text-to-Speech (TTS), and Voice Conversion (VC) techniques that mimic legitimate users to bypass authentication. To address these challenges, this paper proposes a deep learning-based anti-spoofing framework developed and evaluated on the ASVspoof 2019 Physical Access (PA) dataset. Various architectures were examined, including ResNet-based models and hybrid designs that combine convolutional layers with LSTM and BiLSTM units, which are capable of capturing both spatial features and temporal dynamics of speech signals. Results demonstrated that hybrid models consistently outperformed purely convolutional networks, with the ResNet18+BiLSTM model achieving 96.88% accuracy and an Equal Error Rate (EER) of 3.12%, demonstrating strong effectiveness in spoof detection. These findings confirm that combining convolutional feature extraction with temporal modeling greatly improves the overall robustness of anti-spoofing systems.

**Key Words:** Voice Spoofing Detection, Automatic Speaker Verification, Deep Learning, ResNet, ResNet18, LSTM, BiLSTM, Equal Error Rate, ASVspoof 2019.

## 1. INTRODUCTION

Voice-based authentication, commonly known as Automatic Speaker Verification (ASV), has seen growing adoption in domains such as banking, mobile security, and access control, owing to its user-friendly nature and its reliance on the inherent uniqueness of an individual's voice. However, the rapid advancement of technologies including speech synthesis, voice conversion, and deepfake audio generation has introduced significant spoofing threats to these systems. In such attacks, an adversary attempts to replicate or artificially construct a target speaker's voice in order to gain unauthorized access to secured systems. Earlier spoof detection approaches depended largely on manually engineered acoustic features such as MFCC and CQCC. Although these techniques achieved moderate success, they frequently fell short when faced with novel or unseen attack scenarios, making generalization across diverse spoofing methods a persistent limitation. Deep learning has since emerged as a more capable and adaptable alternative.

Residual networks (ResNet) have proven particularly effective at learning discriminative features directly from audio data, while Long Short-Term Memory (LSTM) networks are well-suited for capturing the sequential and temporal characteristics of speech. Nevertheless, relatively little research has directly compared standalone ResNet models against hybrid architectures that merge residual learning with temporal modeling under varying training configurations. To address this research gap, the current study systematically evaluates several model architectures — including ResNet, ResNet+LSTM, ResNet18+LSTM, and ResNet18+BiLSTM — on the ASVspoof 2019 dataset. Models are trained across different epoch settings to assess how training duration and the inclusion of temporal modeling affect spoof detection performance.

## 2. Literature Survey

Wu et al. introduced the ASVspoof 2015 challenge, which established one of the earliest standardized benchmarks for evaluating anti-spoofing countermeasures in Automatic Speaker Verification (ASV) systems. Their framework employed Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) in conjunction with MFCC features to detect replay and synthetic spoofing attacks. Although the system performed adequately against known attack types, it demonstrated notable weaknesses when exposed to previously unseen spoofing scenarios [1].

In a separate investigation, Wu et al. conducted a comprehensive analysis of various spoofing attack categories and their corresponding countermeasures within speaker verification frameworks. Their study concluded that replay, speech synthesis, and voice conversion attacks pose substantial risks to system reliability, reinforcing the critical need for robust anti-spoofing solutions [2].

Todisco et al. presented the ASVspoof 2019 challenge, encompassing both Logical Access (LA) and Physical Access (PA) evaluation conditions. Their work incorporated CQCC, LFCC, and x-vector representations combined with deep neural networks, demonstrating that deep learning-based approaches yield superior accuracy and generalization relative to conventional feature-based methods [3].

Chen et al. addressed the challenge of codec-induced and channel-related variabilities in spoof detection pipelines. They introduced an enhanced ECAPA-TDNN architecture

that incorporates adversarial training strategies and data augmentation techniques, resulting in reduced Equal Error Rates and improved performance under practical deployment conditions [4].

Lei et al. proposed Two-Path GMM-ResNet and GMMSENet architectures that merge statistical modeling with deep residual learning. This combined design enables simultaneous capture of both spectral and statistical speech properties, contributing to more accurate spoof detection outcomes [5].

Kim and Ban developed a phase-aware detection framework utilizing the Res2Net architecture alongside a dedicated network for processing phase information. Their experimental findings indicated that explicitly incorporating phase-related features leads to improved detection accuracy and a reduction in false acceptance rates [6].

Lei et al. subsequently introduced GMM-ResNet2, an ensemble-based framework that integrates multiple ResNet variants with multi-order GMM representations. This architecture demonstrated enhanced generalization capability, particularly when evaluated against previously unseen spoofing attack types [7].

Zhang et al. proposed the AASIST2 model, specifically engineered to handle short-duration speech segments. Through the use of the Res2Net backbone and large-margin fine-tuning strategies, the model maintains consistent and reliable detection performance across varying speech lengths [8].

Wu et al. presented a multiscale feature aggregation framework combined with dynamic convolution operations for anti-spoofing purposes. Their approach exhibited strong detection capability and generalization across multiple benchmark datasets [9].

Rani et al. conducted a comparative study between classical machine learning algorithms, including GMM and SVM, and deep learning architectures such as LSTM. Their results confirmed that LSTM-based models achieve superior detection performance owing to their inherent capacity to model temporal dependencies within speech signals [10].

Boles and Rad proposed a voice authentication framework grounded in MFCC-based feature extraction paired with SVM classification. Their work validated the effectiveness of traditional methodologies while simultaneously acknowledging the growing need for more advanced and robust spoof detection mechanisms [11].

Shahzad et al. constructed a hybrid deep learning model that integrates VGGish, LSTM, YAMNet, and one-dimensional CNN components. The resulting system achieved high classification accuracy and low EER, demonstrating the advantage of fusing diverse deep learning modules within a unified detection pipeline [12].

A study titled "Deep Fake Defender: AI-Based Detection of Deepfake Voice Attacks" presented a detection system targeting synthetically generated and manipulated speech. The framework leverages acoustic feature representations alongside machine learning and CNN-based classifiers to differentiate authentic from fabricated audio, though performance against unseen attack variants remains an open challenge [13].

The work "AI Based Voice Spoofing Detection using ML and DL" systematically evaluated both conventional classifiers such as SVM and Random Forest, and deep learning models including CNN and LSTM architectures. The findings consistently favored deep learning approaches, attributing their superior accuracy to more expressive and generalizable feature representations [14].

A comprehensive review titled "Deepfake Audio Detection in Voice Authentication" surveyed recent advancements in synthetic speech detection. The study concluded that integrating cepstral or spectrogram-derived features with deep learning architectures significantly enhances system robustness against contemporary spoofing techniques [15].

The contribution "DeepLASD Countermeasure for Logical Access Audio Spoofing" introduced a deep learning solution specifically tailored for logical access spoofing scenarios. By employing attention-driven feature learning mechanisms alongside advanced spectral processing strategies, the proposed model achieved competitive benchmark performance [16].

In "Comparative Analysis of ML/DL Models for Voice Spoofing Detection," a diverse set of machine learning and deep learning architectures were assessed under standardized experimental conditions. The study established that deep learning and ensemble-based methods consistently surpass traditional approaches, with particularly notable gains observed during cross-dataset evaluations [17].

The study "Voice Spoofing Countermeasure for Voice Replay Attacks" targeted the detection of replay-based spoofing using acoustically robust feature representations. The proposed methodology demonstrated improved resilience to environmental noise and channel-induced distortions under real-world testing conditions [18].

"DeepDetection: Privacy-Enhanced Deep Voice Detection and Authentication" introduced a unified framework that simultaneously addresses spoof detection and data privacy preservation. The system ensures secure and privacy-conscious audio processing while sustaining high levels of detection accuracy [19].

The paper "Evaluation Framework for Deepfake Speech Detection" established a standardized evaluation methodology for benchmarking anti-spoofing systems. The work emphasized generalization as a core requirement and

highlighted metrics such as EER and Detection Cost Function as essential indicators of system effectiveness [20].

In "Dual-Channel Spoofed Speech Detection Based on Graph Attention Networks," a dual-stream detection model was proposed to jointly capture spectral and temporal speech characteristics through graph attention mechanisms, yielding measurable improvements in overall detection performance [21].

"Voice Spoofing Detector: A Unified Anti-Spoofing Framework" presented a comprehensive system that consolidates multiple acoustic feature streams with deep learning models, demonstrating consistently strong detection results across a range of spoofing conditions [22].

The study "A Blended Framework for Audio Spoof Detection" proposed an ensemble strategy that combines handcrafted acoustic features with learned deep representations. This hybrid integration was shown to enhance system robustness and contribute to a meaningful reduction in false acceptance rates [23].

### 3. Proposed Methodology

#### 3.1 Selection of Dataset

In this study, the Physical Access (PA) subset of the ASVspoof2019 dataset was selected as the primary experimental resource. This dataset holds significant relevance as it encompasses both authentic speech recordings and replay-based spoofed audio captured under realistic acoustic environments. Adhering to the standard evaluation protocol, the dataset was partitioned into three distinct subsets: training, development, and evaluation. To mitigate the effects of class imbalance, balanced sampling was applied during preparation of the training and development subsets, ensuring proportional representation of both genuine and spoofed samples. The evaluation subset was maintained entirely separate throughout the experimental process to guarantee unbiased and objective performance assessment.

#### 3.2 Preprocessing

Prior to feature extraction, all audio recordings were subjected to a uniform preprocessing pipeline to ensure consistency across the dataset. The applied steps included:

- Silence removal to discard non-informative audio segments
- Amplitude normalization to achieve uniform signal scaling
- Resampling to a standardized rate of 16 kHz
- Conversion to single-channel mono format

These preprocessing operations collectively reduce variability introduced by differing recording environments and background noise conditions, thereby enabling more stable and reliable downstream feature extraction.

#### 3.3 Feature Extraction

Constant-Q Cepstral Coefficients (CQCC) were extracted from the preprocessed audio signals as the primary feature representation. CQCC features are especially well-suited for replay attack detection due to their ability to capture fine-grained frequency information, with particular sensitivity in lower frequency regions. The extracted feature representations were stored as two-dimensional NumPy arrays to facilitate efficient batch processing during model training. In essence, CQCC equips the model with the capacity to identify subtle acoustic distinctions that differentiate genuine speech from spoofed recordings

#### 3.4 Model Architecture

To investigate the impact of integrating spatial and temporal learning mechanisms, four distinct architectural configurations were designed and evaluated:

- ResNet
- ResNet + LSTM
- ResNet18 +LSTM
- ResNet18 + Bidirectional LSTM (Proposed Model) In the proposed architecture, ResNet18 serves as the backbone network responsible for extracting discriminative spectral features from the CQCC inputs. Its residual connections facilitate more effective gradient flow during training, alleviating the vanishing gradient problem. The resulting feature representations are subsequently passed to a Bidirectional LSTM layer, which models temporal dependencies across both forward and backward time directions, thereby enriching the model's capacity to understand the dynamic nature of speech signals.

#### 3.5 Training Strategy

All model configurations were optimized using the cross-entropy loss function in combination with the Adam optimizer. Dropout layers were incorporated throughout the networks to mitigate overfitting, and early stopping was employed to terminate training whenever validation performance ceased to improve. Each model was trained across varying epoch counts to analyze differences in learning behavior and to assess generalization capability under different training durations.

#### 3.6 Evaluation metrics

Model performance was assessed on the PA evaluation subset using the following metrics:

- Accuracy
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC)
- Equal Error Rate (EER)
- False Acceptance Rate (FAR)
- False Rejection Rate (FRR)

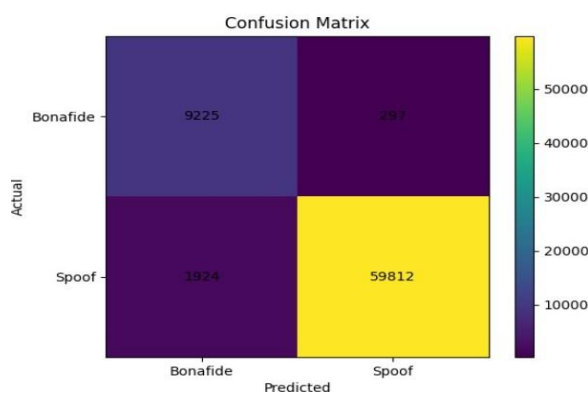
The Equal Error Rate is calculated at the point where FAR and FRR are equal. In addition, confusion matrices were used to better understand how well the model distinguishes between genuine and spoofed samples.

### 3.7 Deployment Framework

During inference, incoming audio inputs are processed through the same preprocessing and CQCC feature extraction pipeline used during model training, ensuring consistency between training and deployment conditions. The resulting feature representations are then forwarded to the trained ResNet18 + BiLSTM model to generate predictions. The classification threshold is calibrated based on EER analysis to strike an appropriate balance between security stringency and system usability. The system ultimately produces a binary classification output indicating whether the submitted speech sample is authentic or spoofed.

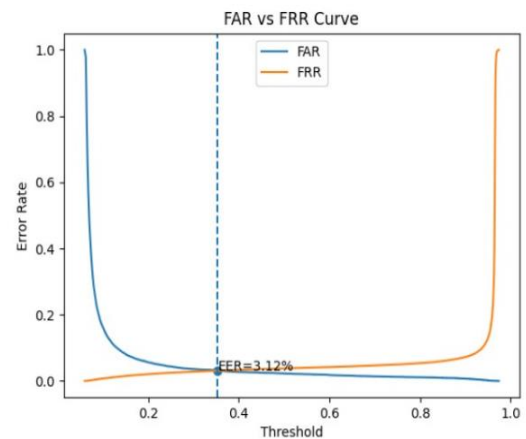
## 4. Results and Description

### 4.1 ResNet18+BiLSTM performance at 13 epochs



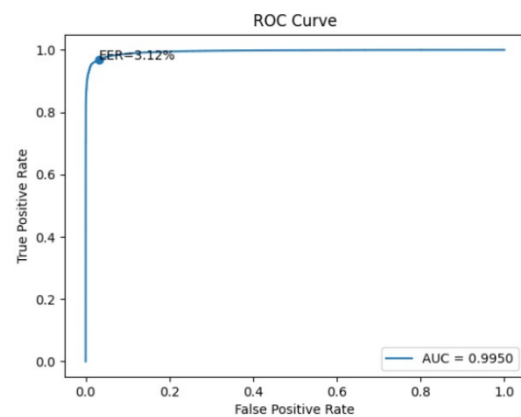
**Fig - 1:** Confusion matrix of ResNet18+BiLSTM at 13 epochs.

The matrix confirms precise identification of 9225 spoofed and 59812 bonafide samples, with 297 false negatives and 1924 false positives documented during evaluation.



**Fig - 2:** FAR vs FRR of ResNet18+BiLSTM at 13 epochs.

The graph reveals the reciprocal behavioral pattern of FAR and FRR as the decision threshold progressively varies. The specifically marked point of convergence between the two curves signifies the Equal Error Rate (EER), which acts as the fundamental criterion for determining overall model performance.



**Fig - 3:** ROC curve of ResNet18+BiLSTM at 13 epochs.

The model records an AUC value of 0.9950, establishing exceptional predictive accuracy with consistently high sensitivity levels and a distinctly reduced rate of false positive classifications.

## 4.2 Performance comparison across all models

**Table -1:** Comparison of performance metrics across different model architectures and training configurations

Metrics	ResNet at 5 Epochs	ResNet at 10 Epochs	ResNet +LSTM at 5 Epochs	ResNet +LSTM at 10 Epochs	ResNet 18 +LSTM at 10 Epochs	ResNet18+ BiLSTM at 12 epochs	ResNet18+ BiLSTM at 13 epochs
Accuracy	73.06%	77.03%	88.44%	89.38 %	88.59%	96.29%	96.88%
AUC-ROC	0.796	0.8372	0.9472	0.9544	0.9572	0.9939	0.99500
EER	27.20%	23.64%	12.78%	11.42%	11.10%	3.71%	3.12%
FAR	41.18%	28.19%	12.30%	11.41%	9.57%	3.71%	3.12%
FRR	15.23%	18.68%	12.39%	11.43%	13.29%	3.71%	3.12%
TN	5198	6346	4340	4454	8775	8500	59812
FP	3639	2491	1060	946	929	1195	297
FN	1637	2007	189	201	1265	1800	1924
TP	9109	8739	5211	5199	8257	9000	9225

Table -1 presents a comparative summary of Accuracy, AUC-ROC, EER, FAR, and FRR across all evaluated model architectures. The proposed ResNet18+BiLSTM model attains the highest accuracy and AUC-ROC values while simultaneously recording the lowest error rates among all configurations, confirming its superior robustness and overall effectiveness in detecting replay-based spoofing attacks

## 5. Conclusion

This study proposed a deep learning-driven framework for spoofed speech detection, developed and assessed using the Physical Access (PA) subset of the ASVspoof 2019 dataset. The central objective was to determine how effectively deep learning architectures can differentiate between authentic speech and replay-based spoofed audio under realistic acoustic conditions. Experimental outcomes confirmed that convolutional networks demonstrate strong capability in capturing meaningful acoustic representations, while the additional integration of temporal modeling components further elevates overall detection performance. Of all the architectural configurations evaluated, the ResNet18+BiLSTM model produced the most consistent results, attaining a lower Equal Error Rate than all competing architectures, thereby demonstrating the clear advantage of jointly leveraging spatial feature extraction and temporal sequence modeling. Collectively, the results suggest that deep learning-based approaches hold strong potential for strengthening the robustness and dependability of voice-based authentication systems in the face of replay spoofing threats. Furthermore, this work establishes a solid foundation for future investigations, wherein a broader range of spoofing attack categories and more extensive datasets may be explored to push detection performance to greater heights.

## REFERENCES

- [1] Z. Wu et al., "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in Proc. INTERSPEECH, Dresden, Germany, 2015, pp. 2037–2041.
- [2] Z. Wu et al., "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, vol. 66, pp. 130–153, 2015.
- [3] M. Todisco et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in Proc. INTERSPEECH, Graz, Austria, 2019, pp. 1008–1012.
- [4] J. Chen et al., "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 29, pp. 2372–2385, 2021.
- [5] Z. Lei et al., "Two-path GMM-ResNet and GMM-SENet for ASV spoofing detection," in Proc. IEEE ICASSP, Singapore, 2022, pp. 6377–6381.
- [6] J. Kim and S. M. Ban, "Phase-aware spoof speech detection based on Res2Net with phase network," in Proc. IEEE ICASSP, Rhodes Island, Greece, 2023, pp. 1–5.
- [7] Z. Lei et al., "GMM-ResNet2: Ensemble of group ResNet networks for synthetic speech detection," in Proc. IEEE ICASSP, Seoul, South Korea, 2024, pp. 12101–12105.
- [8] Y. Zhang et al., "Improving short utterance anti-spoofing with AASIST2," in Proc. IEEE ICASSP, Seoul, South Korea, 2024, pp. 11636–11640.
- [9] H. Wu et al., "Robust spoof speech detection based on multi-scale feature aggregation and dynamic convolution," in Proc. IEEE ICASSP, Seoul, South Korea, 2024, pp. 10156–10160.
- [10] N. Rani et al., "A review on machine learning approaches for deepfake voice detection," Journal of Intelligent Systems, vol. 33, no. 1, 2024.
- [11] W. Boles and L. Rad, "Voiceprint authentication using deep learning and MFCC features," IEEE Access, vol. 5, pp. 17112–17120, 2017.
- [12] A. Shahzad et al., "VGGish-LSTM hybrid model with YAMNet and 1D CNN for spoof speech detection," in Proc. IEEE ICASSP, 2025, pp. 1–5.
- [13] F. M. Fazeeha et al., "Deep fake defender: AI-based detection of deepfake voice attacks in real-time voice authentication systems," Int. J. Res. Appl. Sci. Eng. Technol., vol. 13, no. 8, pp. 1061–1064, 2025.
- [14] M. Darshan and H. N. Poornima, "AI based voice spoofing detection using machine learning and deep

learning,” *Int. J. Sci. Comput. Informatics*, 2023.  
[Online]. Available: <https://ijsci.com/index.php/home/article/view/983>

[15] A. O. M. Salih et al., “Deepfake audio detection in voice authentication: A review,” *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 1, pp. 14390–14395, 2024. [Online]. Available: <https://etasr.com/index.php/ETASR/article/view/13400>

[16] H. Al-Tairi et al., “DeepLASD countermeasure for logical access audio spoofing,” *Scientific Reports*, vol. 15, 2025.

[17] R. Rani and B. Kishan, “Comparative analysis of machine learning and deep learning models for voice spoofing detection,” *Int. J. Bus. Appl. Sci.*, 2023. [Online]. Available: <https://www.sciencepubco.com/index.php/IJBAS/article/view/34940>

[18] J. Zhou et al., “Voice spoofing countermeasure for voice replay attacks,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2022, no. 1, 2022.

[19] Y. Kang et al., “DeepDetection: Privacy-enhanced deep voice detection and authentication,” *Applied Sciences*, vol. 12, no. 3, 2022.

[20] A. Firc et al., “Evaluation framework for deepfake speech detection,” *EURASIP J. Information Security*, vol. 2024, no. 1, 2024. [Online]. Available: <https://link.springer.com/article/10.1186/s42400-024-00346-1>

[21] Y. Tan et al., “Dual-channel spoofed speech detection based on graph attention networks,” *Symmetry*, vol. 17, no. 5, 2025.

[22] A. Javed et al., “Voice spoofing detector: A unified anti-spoofing framework,” *Expert Systems with Applications*, vol. 195, 2022.

[23] M. Sharafudeen et al., “A blended framework for audio spoof detection,” *Scientific Reports*, vol. 14, 2024.