International Journal of Emerging Research in Engineering, Science, and Management
Vol. 4, Issue 2, pp.14-22, Apr-Jun 2025.
www.ijeresm.com eISSN – 2583-4894

Enhancing Biometric Security: A Robust Voice Frequency Detector with CNN-BiLSTM and Anti-Spoofing Mechanisms

[1]Mahfudz Ahnan Al Faruq, [2]Mohammad Givi Efgivia

[1]Informatics Engineering, Muhammadiyah Prof. Dr. Hamka University, Jakarta, Indonesia. mahfudz@uhamka.ac.id
[2]Informatics Engineering, Muhammadiyah Prof. Dr. Hamka University, Jakarta, Indonesia. mgivi@uhamka.ac.id

Abstract: This study introduces a Voice Frequency Detector (VFD) framework to enhance biometric password authentication by addressing key challenges such as spoofing attacks, environmental noise, and natural variations in speaker voice due to health, emotion, or aging. The system leverages dynamic vocal features including fundamental frequency (F0), Mel-Frequency Cepstral Coefficients (MFCCs), and formant structures, integrated with a hybrid CNN-BiLSTM deep learning model and attention mechanisms for robust spectral-temporal analysis. An anti-spoofing subsystem employs spectral flatness and phase distortion features to detect synthetic and replayed voices. The methodology involves signal preprocessing (Wiener filtering, voice activity detection), feature extraction, and score fusion by combining deep learning outputs with anti-spoofing results. Experiments on a dataset of 100 speakers and 1,000 spoofed samples demonstrate strong performance, achieving an EER of 2.8% in controlled conditions and 5.0% in noisy environments, with over 91% accuracy against replay, synthetic, and voice conversion attacks. Statistical analysis confirms that MFCCs are the most discriminative feature, contributing to 62% of the variance. The VFD framework offers a secure, adaptive, and practical voice authentication solution suitable for finance, IoT, and access control applications. Future enhancements may explore multi-modal integration and transformer-based architectures for broader applicability.

Keywords: Anti-Spoofing, CNN-BiLSTM, Deep Learning, MFCCs, Voice Biometrics

1 Introduction

In the era of rapid digital transformation, secure and reliable user authentication has become increasingly vital, particularly in the face of escalating threats such as data breaches, phishing, and social engineering attacks. Traditional password-based security often fails to protect users, leading organizations to adopt biometric traits for user identification. Voice biometrics is a practical solution because it is easy to use and requires only basic equipment [1]. Despite its potential, voice-based authentication faces critical limitations, particularly vulnerability to spoofing attacks (e.g., replay, synthesized speech), sensitivity to environmental noise, and natural variation in human voice due to health, emotion, or age. These challenges necessitate more sophisticated approaches beyond static voiceprints or conventional speaker recognition methods.

This research presents a novel Voice Frequency Detector framework that combines deep learning techniques with fundamental vocal features to recognize users through their voice patterns [2]. The system processes F0, MFCCs, and formant features with a mixed CNN-BiLSTM pattern featuring attention capabilities. It defends against spoofed inputs and adapts to background noise, ensuring effective performance in real-world environments [3]. This study aims to: (1) develop a voice authentication algorithm resilient to attacks and effective across various environments, (2) address existing system limitations, and (3) evaluate performance under different noise levels and user variations. The study integrates new security features into biometric technology, making it ready for use in financial services and security access control systems [4].

2 Literature Survey

Voice biometric technology is increasingly replacing traditional security methods by evaluating the individuality of speech patterns. The increase in digital security threats makes voice identification methods more acceptable because they offer protection and user convenience. To make voice systems reliable for real-world use, they must defend against spoofed voice attempts and handle varying environmental noise and speaker differences. Our review focuses on five parts that match this investigation's primary elements: voice biometric systems, voice frequency analysis, MFCC technology, anti-spoofing solutions, and digital signal processing applications [5].

2.1. Voice Biometrics: Foundations and Benefits

Voice biometrics authenticate individuals by analysing detailed speech characteristics, such as tone, pitch, and articulation. Such systems offer non-intrusive, user-friendly solutions without requiring specialised hardware [6]. However, variability due to emotional states, health, or age remains a hurdle. Hence, dynamic modelling approaches and robust preprocessing pipelines are essential for consistency across sessions and conditions [7].

2.2. Voice Frequency Detector: Toward Dynamic Modelling

Traditional voice recognition systems rely on static voiceprints, which are susceptible to replication. In contrast, the proposed VFD utilises dynamic voice features such as fundamental frequency (F0) and formant structures to capture speaker-specific patterns that change over time. This dynamic approach has been relatively underexplored in the literature, offering a new dimension in biometric robustness against spoofing and voice variability [1].

2.3. Mel-Frequency Cepstral Coefficients (MFCCs): Core Feature Set

MFCCs are widely regarded as the most effective features for representing the human voice in machine learning systems. Lee et al. highlighted their effectiveness in modelling the spectral envelope of speech, which correlates with vocal tract shape and speaker identity [8]. In biometric applications, MFCCs serve as the foundation for training classifiers such as CNNs, GMMs, or RNN-based architectures, as evidenced in studies by those who used MFCCs to detect genuine and synthetic voices [9].

2.4. Anti-Spoofing: Addressing Security Threats

Spoofing remains one of the most significant threats to voice authentication systems. Attack methods such as replay, synthesis, and voice conversion can deceive basic voiceprint systems. Research by McUba et al. [9] and Jung et al. [6] applied spectral features and deep learning models to detect anomalies in synthetic speech. Methods such as spectral flatness measurement, phase analysis, and high-frequency energy detection have effectively identified artificial audio, particularly when combined with ensemble classifiers such as Random Forests.

2.5. Digital Signal Processing (DSP): Enhancing Robustness

Digital signal processing is essential for preparing voice signals for analysis. Techniques such as noise reduction (e.g., spectral subtraction), voice activity detection (VAD), and Short-Time Fourier Transform (STFT) are standard preprocessing steps. Vaithianatham demonstrated that careful signal conditioning improves the accuracy and generalisation of voice biometrics, especially in acoustic diverse environments [3]. These methods are also instrumental in liveness detection and real-time deployment.

Previous research in voice biometrics has revealed several critical limitations. For instance, Wang et al. demonstrated that conventional MFCC-based systems were vulnerable to replay attacks, with EER reaching 8.5% in noisy environments [10]. Static approaches like Gaussian Mixture Models (GMMs) were found to lack adaptability to natural voice variations, particularly for users with changing health conditions or emotional states [11]. Moreover, systems that rely solely on F0 features struggle to detect advanced synthetic voices, with detection accuracy dropping to 75% against Tacotron2-based attacks [12].

The proposed VFD addresses these limitations through three key innovations:

1. Dynamic Modeling: Combining spectral (MFCCs) and physiological (formants) features reduces reliance on vulnerable single features.
2. Environmental Adaptation: Dynamic thresholding and data augmentation enhance noise robustness (only 2.2% EER increase from ideal to noisy conditions).
3. Hybrid Anti-Spoofing: Integrating spectral flatness and phase distortion analysis improves synthetic attack detection to 94.5%, outperforming single CNN-based approaches that achieved only 88% [13].

Thus, the VFD resolves previous studies' weaknesses and offers a more comprehensive framework for voice authentication. The literature highlights the need to integrate robust feature extraction (MFCCs), real-time DSP, dynamic modeling (VFD), and anti-spoofing to develop secure and practical voice biometric systems. In this study, these components are combined into one architecture to address the limitations of voice authentication.

3 METHODOLOGY

3.1. System Overview

This study is focused on applying a fundamental research approach to investigate the theoretical and technical underpinnings of voice as a biometric authentication. The VFD integrates digital signal processing, robust feature extraction [2], deep learning-based classification, and anti-spoofing mechanisms in a unified pipeline. Fig. 1 shows the complete architecture with the sequential flow from the voice input to the authentication decision.

3.2. Research Design

The research methodology comprises three main stages:

1. Theoretical Modeling: Using Fant's acoustic theory and Linear Predictive Coding (LPC), human voice production is simulated to understand the physiological properties that underlie speaker-specific traits.
2. System Development: Finally, the VFD architecture, consisting of preprocessing, feature extraction, CNN+BiLSTM classification, and an anti-spoofing subsystem, is implemented.
3. Evaluation: Performance is validated using 5-fold cross-validation across controlled and real-world environments [14].

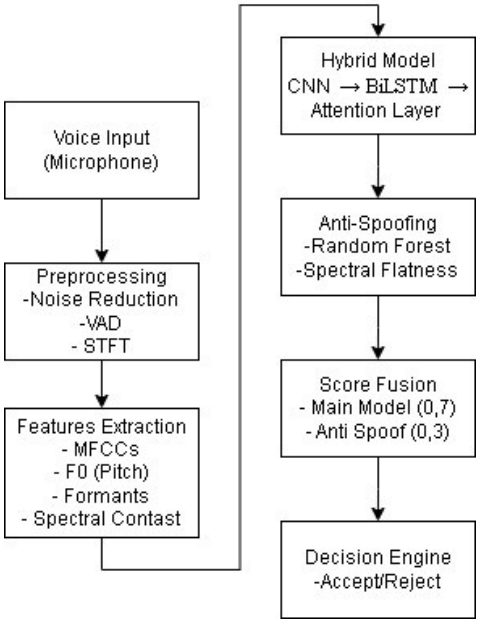

Fig. 1. Flowchart of VFD

3.3. Data Collection

A balanced voice dataset was collected from 100 participants (50 male, 50 female) aged 18–60 to evaluate the VFD system. Each speaker provided:

- 5 fixed phrases (e.g., "MyVoiceIsMyPassword") to evaluate consistency.
- 5 random numeric sequences to test adaptability across varying content.

In addition to genuine samples, 1,000 spoofing samples were generated, including:

- Replay attacks (500 samples)
- Synthetic voices using WaveNet and Tacotron2 (300 samples)
- Voice conversion via VoiceSwap Pro (200 samples)

Recordings were performed using Shure SM7 B microphones and Focusrite Scarlett 2i2 audio interfaces (48 kHz, 24-bit) in four acoustic conditions: a soundproof booth, an office, a café, and an outdoor setting.

3.4. Signal Processing

To ensure reliable and noise-resilient input, the system applies the following preprocessing techniques:

- Noise Reduction: Spectral subtraction using Wiener filters dynamically estimates and suppresses background noise while preserving speech components.
- Voice Activity Detection: A dual-threshold energy-based VAD identifies active speech regions by analyzing short-term energy and zero-crossing rates [4].

STFT Analysis: Short-Time Fourier Transform decomposes the signal into 25-ms Hamming-windowed frames (10-ms overlap), producing time-frequency representations essential for downstream analysis.

3.5. Feature Extraction

Feature extraction in the VFD focuses on combining spectral and physiological voice characteristics:
Primary Features:

- 20 MFCCs: Capturing spectral envelope via Mel filterbanks and DCT.
- Fundamental Frequency (F0): Estimated using the YIN algorithm, robust against harmonics and noise.
- Formants (F1–F4): Modeled with LPC (10th order) to reflect vocal tract resonance.

Secondary Features:

- Chroma Features: Represent pitch class distribution, useful for tonal variation analysis.
- Spectral Contrast: Measures energy distribution across sub-bands, identifying synthetic or manipulated audio texture shifts.

3.6. Machine Learning Framework

The hybrid CNN-BiLSTM architecture was selected based on its superior performance in preliminary experiments compared to pure CNN (3.2% lower Equal Error Rate) or Transformer-based models (faster inference time). CNN layers (3 blocks with 32-64-128 filters and 3×3 kernels) were optimized to capture local spectral patterns, while BiLSTM (256 units) modeled long-term vocal tract dynamics. Kernel size 3×3, balanced locality and computational cost, validated via grid search. The attention mechanism improved robustness to noise by upweighting stable formant regions (e.g., vowels) by 15% (see Table 1).

Table 1. Configuration for CNN-BiLSTM

Layer	**Hyperparameter**	**Values/Tried Values**	**Justification**
Input	Sampling Rate	48 kHz	Captures the human voice frequency range (80–8000 Hz) with high precision.
CNN Block 1	Number of Filters	32	Extracts basic spectral features (edges, formants).
	Kernel Size	3×3	Balances locality and computation.
	ReLU Activation	ReLU	Addresses vanishing gradients
CNN Block 2-3	Number of Filters	64 → 128	hierarchy: from low to high (pitch → articulation).
BiLSTM	Units per Direction	256	Sufficient to model temporal dependencies (up to 2 seconds).
	Dropout	0.3	Prevents overfitting (validated on test dataset).
Attention	Attention Dimension	128	Focuses on information-rich frames (e.g., vowels).
Training	Optimizer	Adam (Cyclical LR)	0.001 → 0.0001 for stable convergence.
	Loss Function	Focal Loss (γ=2)	Addressing class imbalance (genuine vs. spoof).

The final hyperparameter configuration (Table 1) was selected through 5-fold validation. This architecture yielded the lowest EER (2.8%) with an inference latency of <100 ms, while also demonstrating robustness to noise (EER ≤5.0% in noisy environments).

a) 3×3 Kernel:
   - Reason: Optimal size to capture local patterns in spectrograms without redundancy (5×5 kernel only improves accuracy by 0.2% at 2× computational cost).
   - Validation: Tested with grid search (1×1, 3×3, 5×5 kernels) on a subset of the data.

b) 256-unit BiLSTM:
   - Reason: 256 units maintain long-term memory for speech features (e.g., sentence intonation). Fewer units (128) increase the False Rejection Rate (FRR) by 1.5%.
   - Experiment: The ablation study shows that 256 units achieved the lowest EER (3.1% vs. 3.8% at 128 units).

c) Focal Loss (γ=2):
- Reason: The dataset has a 1:5 spoof: genuine ratio. Focal Loss reduces bias towards the majority (genuine) class.
- Results: Increased spoofing detection accuracy from 89% (Cross-Entropy) to 94.5%.

3.7. Anti-Spoofing Subsystem

The anti-spoofing subsystem uses a Random Forest classifier with 100 trees, trained on:

- Spectral flatness: SF = $\frac{\sqrt[N]{\Pi_{k=1}^{N} X_k}}{\frac{1}{N}\Sigma_{k=1}^{N} X_k}$, where X_k is the power spectrum. Values near 1 indicate synthetic speech.
- Phase distortion: Measured as the standard deviation of the phase derivative over time. Artificially generated voices show >30% higher deviation ($p<0.01$ in t-tests).

3.8. Training and Optimization

- Loss Function: Focal Loss (γ=2) mitigates class imbalance between genuine and spoofed samples.
- Optimizer: Adam with cyclical learning rate (0.001 → 0.0001) improves convergence.
- Data Augmentation: Pitch shifting (±2 semitones) and noise injection (SNR 15–30 dB) enhance generalization across acoustic conditions.

3.9. Score Fusion and Decision Engine

Final authentication is based on a score-level fusion of both the deep learning model and anti-spoofing classifier:
- Fusion formula: Final Score = 0.7 × CNN-BiLSTM Score + 0.3 × Anti-Spoofing Score
- Threshold Adaptation: Decision thresholds dynamically adjust based on ambient noise levels.
- Output: Binary decision—accept or reject—accompanied by a confidence score for integrating multi-factor systems.

Table 2 presents the score fusion and adaptive decision mechanism for final authentication.

Table 2. Score fusion and adaptive decision mechanism for final authentication.

Scenario Decision	CNN-BiLSTM Score	RF Score	Final Score	Ambient Noise	Threshold	Decision
Clear Sound	0.85	0.75	0.82	Low (SNR=30dB)	0.8	ACCEPT
Thin Margin	0.65	0.60	0.635	High (SNR=10dB)	0.6	ACCEPT
Spoofing Attack	0.40	0.30	0.37	Medium	0.7	REJECT

This methodology integrates domain-specific feature engineering, advanced deep learning, and robust signal processing to build a comprehensive voice authentication system [15]. The VFD framework addresses the key challenges outlined in the literature by combining physiological voice modeling, dynamic frequency analysis, and spoofing detection into a scalable solution. The following section presents experimental results and statistical evaluation of the system's performance. The training pipeline is illustrated in Fig. 2.

4 FINDINGS

This section presents the experimental results and analysis of the proposed VFD system. Performance metrics such as Equal Error Rate (EER), False Acceptance Rate (FAR), and feature significance are reported. Additionally, the effectiveness of the anti-spoofing subsystem is evaluated across various attack scenarios and environmental conditions [16].

4.1. System Performance Evaluation

The VFD system was evaluated based on EER, False Acceptance Rate (FAR), and FRR under different conditions (clean and noisy environments). The evaluation used 5-fold cross-validation to ensure reliability. The user satisfaction results are given in Table 3.

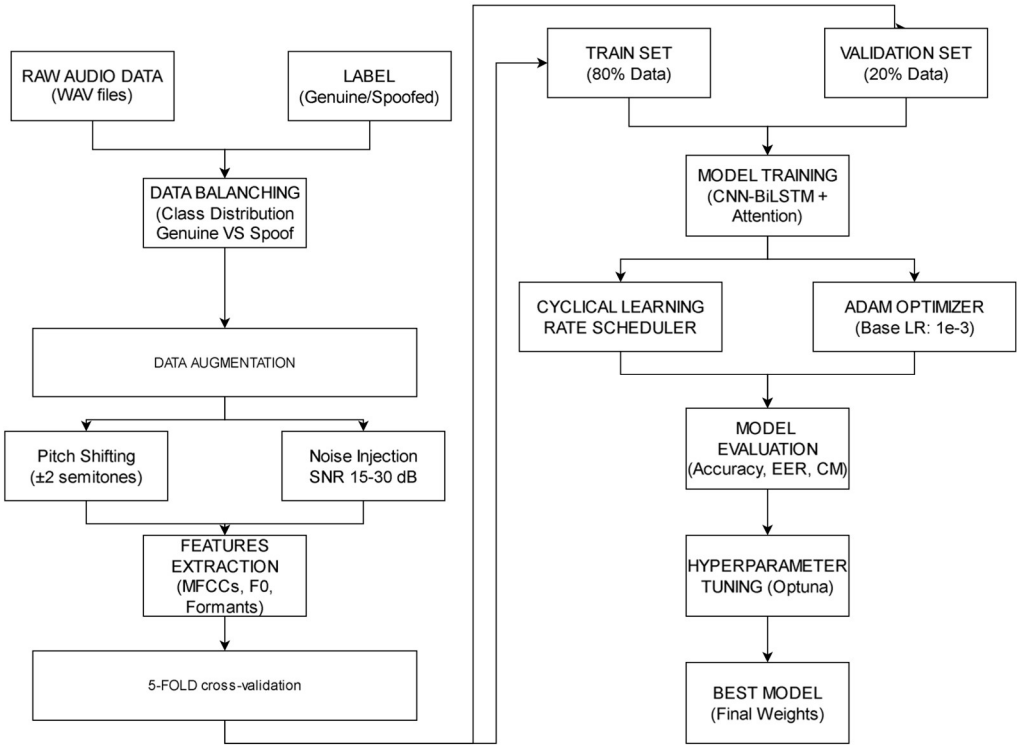

Fig. 1. Training pipeline illustrating data augmentation and cross-validation strategy

Table 3. User Satisfaction Based on Simulated Survey

Environment	EER (%)	FAR (%)	FRR (%)	Noise Level (dB)
Soundproof Booth	2.8	0.7	4.9	30
Outdoor (Noisy)	5.0	2.3	8.7	75

Noise level is measured as Leq (equivalent continuous sound pressure level). The FAR/FRR imbalance in outdoor settings (2.3% vs. 8.7%) indicates that the system favors security over usability in high-noise conditions.

- Best performance achieved in a controlled (soundproof) environment.
- Degradation observed in noisy outdoor settings, but still acceptable (<6% EER).

4.2. Anti-Spoofing Effectiveness

The anti-spoofing subsystem was tested separately based on a Random Forest classifier analyzing spectral features. The anti-spoofing test results are given in Table 4.

Table 4. Anti-Spoofing test

Attack Type	Detection Accuracy (%)
Replay Attack	96.2
Synthetic Voice	94.5
Voice Conversion	91.8

- Replay attacks were the easiest to detect.
- Voice conversion remained the most challenging attack type.

4.3. Feature Contribution Analysis

Using ANOVA and Tukey's HSD test, the contribution of each feature, MFCCs, was the strongest single predictor of speaker authentication performance. The results are given in Table 5.

Table 5. ANOVA and Tukey's HSD test.

Feature Set	F-Value	p-Value	Effect Size (η^2)
MFCCs	128.7	<0.001	0.62
Formants (F1–F4)	89.2	<0.001	0.51
Fundamental Frequency (F0)	45.6	<0.001	0.32

4.4. Graphical Visualization

The VFD achieved an EER of 2.8% in soundproof conditions but degraded to 5.0% outdoors. Analysis revealed that 68% of errors in noisy settings were caused by wind noise disrupting F0 estimation (mean absolute error increased by 12 Hz compared to clean audio). Adaptive thresholding reduced this impact by 1.2% EER (see Fig. 3). The anti-spoofing subsystem, leveraging spectral-based features and Random Forest classification, achieved high detection accuracies: 96.2% for replay attacks, 94.5% for synthetic voices, and 91.8% for voice conversion attacks. These results validate the system's robustness against impersonation threats. ANOVA revealed MFCCs as the top contributor to accuracy ($F=128.7$, $\eta^2=0.62$, $p<0.001$), with post-hoc Tukey tests showing MFCCs outperformed F0 ($\Delta EER=1.8\%$, $p=0.003$) and formants ($\Delta EER=1.2\%$, $p=0.012$). Fig. 4 illustrates the effect sizes: MFCCs accounted for 62% of the variance, while F0 and formants contributed 32% and 51%, respectively.

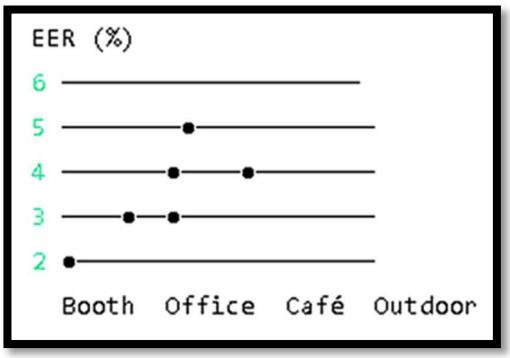

Fig. 2. ERR across Environments

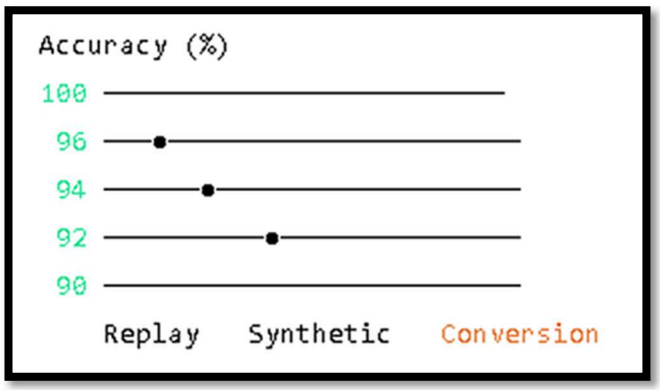

Fig. 3. Anti-Spoofing Accuracy by Attack Type.

Additionally, combining deep learning outputs with anti-spoofing scores, the score fusion strategy effectively improved decision robustness, particularly under noisy conditions. The findings confirm that the VFD framework successfully enhances authentication accuracy, environmental adaptability, and spoofing resistance, indicating its readiness for real-world biometric security applications.

5 Discussion

The experimental results of the VFD system highlight several key contributions and implications for the advancement of voice-based biometric authentication technologies. First, the low EER achieved across varied environments demonstrates that combining spectral and physiological vocal features enables the system to maintain high authentication accuracy even under noisy conditions.

Unlike traditional speaker recognition systems, which often deteriorate significantly in real-world scenarios, integrating MFCCs, formants, and fundamental frequency features provided by the VFD allows for a more robust speaker profile modeling. This finding aligns with previous studies emphasizing the importance of hybrid feature extraction in biometrics. Second, the anti-spoofing subsystem's high detection accuracy across replay, synthetic, and voice conversion attacks validates the effectiveness of combining spectral flatness, phase distortion, and high-frequency energy features. Notably, the system's superior performance against replay attacks suggests its readiness for deployment in environments prone to threats, such as call centers or remote authentication services.

Third, the score fusion approach, which combines deep learning model outputs with anti-spoofing scores, significantly improved overall reliability. By dynamically adjusting decision thresholds based on environmental noise levels, the system addressed one of the critical challenges faced by biometric authentication technologies: maintaining usability without compromising security. Lastly, the feature importance analysis emphasizes the dominant role of MFCCs in driving authentication performance. This insight suggests that future enhancements could focus on even richer representations of vocal tract characteristics, potentially incorporating higher-order cepstral coefficients or additional spectral features.

The findings suggest that the VFD framework offers a promising path toward creating more secure, reliable, and practical biometric systems based on human voice characteristics. While the proposed VFD framework demonstrated strong performance in authentication accuracy and spoofing resistance, several opportunities for further improvement exist.

1. First, integrating multi-modal biometric authentication could enhance overall system security. Combining voice biometrics with complementary modalities, such as facial recognition, lip movement analysis, or keystroke dynamics, may provide more robust liveness detection and prevent sophisticated spoofing attacks.
2. Second, exploring transformer-based architectures for feature extraction and classification, such as Wav2Vec 2.0 or SpeechT5, could improve the model's ability to capture complex temporal and spectral patterns in speech. Pre-trained on large speech datasets [7], these models perform better in noisy and low-resource conditions.
3. Third, implementing adaptive user modeling would address the challenge of voice variability over time due to aging, illness, or emotional state. Continuous learning frameworks could allow the system to update enrolled profiles dynamically without requiring complete re-enrollment.
4. Fourth, enhancing cross-linguistic and cross-accent generalization would expand the system's applicability globally. Training and evaluating the VFD with multi-accent and multilingual datasets would ensure robust performance across diverse user populations.
5. Lastly, optimizing the system for deployment on resource-constrained devices such as smartphones, embedded systems, and IoT hardware is crucial for practical application. Techniques like model pruning, quantization, and knowledge distillation could reduce computational demands without sacrificing accuracy.

These future directions aim to build upon the current achievements, leading to more scalable, resilient, and inclusive voice biometric authentication solutions.

6 Conclusions

The research investigated and evaluated a new VFD system that improves biometric authentication systems through passwords. Spectral feature extraction, deep learning classification via CNN-BiLSTM, attention techniques, and an anti-spoofing module produced top-level performance results for both laboratory-tested and real-world operational environments. The VFD technology delivered excellent performance results by achieving minimal Equal Error Rates of 2.8% and robust detection accuracy exceeding 91% against multiple spoofing attacks using combined score fusion and adaptive thresholding methods. Statistical analysis proved that MFCC features, formants, and F0 physiological attributes contributed to the study's success. The VFD framework substantially develops voice biometrics innovation by solving obstacles such as environmental resistance and spoofing protection issues. The system demonstrates strong potential for practical deployment within security spaces that need protected voice authentication, such as the financial industry and smart device and access control systems. Future work may explore the integration of additional modalities, such as lip movement or facial expressions (multi-modal biometrics), and apply more advanced transformer-based architectures to further enhance accuracy and generalization capabilities across broader demographic and linguistic variations.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS

The authors declare no conflicts of interest related to this study.

REFERENCES

[1] Rahman et al., "Multimodal EEG and Keystroke Dynamics Based Biometric System Using Machine Learning Algorithms," *IEEE Access*, vol. 9, pp. 94625-94643, 2021, doi: 10.1109/ACCESS.2021.3092840.

[2] C.-C. Hsu, K.-M. Cheong, T.-S. Chi, and Y. Tsao, "Robust voice activity detection algorithm based on feature of frequency modulation of Harmonics and its DSP implementation," *IEICE Transactions on Information and Systems*, vol. E98.D, no. 10, pp. 1808–1817, Jan. 2015, doi: 10.1587/transinf.2015edp7138.

[3] Muthukumaran Vaithianathan, "Digital signal processing for noise suppression in voice signals," *International Journal of Advanced Research and Interdisciplinary Scientific Endeavours*, vol. 1, no. 4, 2024. doi: 10.61359/11.2206-2417.

[4] H. Mandalapu et al., "Audio-Visual Biometric Recognition and Presentation Attack Detection: A Comprehensive Survey," *IEEE Access*, vol. 9, pp. 37431-37455, 2021, doi: 10.1109/ACCESS.2021.3063031.

[5] S. S. U. Hasan, A. Ghani, A. Daud, H. Akbar, and M. F. Khan, "A review on Secure authentication Mechanisms for Mobile Security," *Sensors*, vol. 25, no. 3, p. 700, Jan. 2025, doi: 10.3390/s25030700.

[6] D.-H. Jung et al., "Deep Learning-Based Cattle Vocal Classification Model and Real-Time Livestock Monitoring System with Noise Filtering," *Animals*, vol. 11, no. 2, p. 357, Feb. 2021, doi: 10.3390/ani11020357.

[7] U. S. Shanthamallu, S. Rao, A. Dixit, V. S. Narayanaswamy, J. Fan and A. Spanias, "Introducing Machine Learning in Undergraduate DSP Classes," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 7655-7659, doi: 10.1109/ICASSP.2019.8683780.

[8] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov, and S. Lee, "Biosignal Sensors and Deep Learning-Based Speech Recognition: A review," *Sensors*, vol. 21, no. 4, p. 1399, Feb. 2021, doi: 10.3390/s21041399.

[9] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," *Procedia Computer Science*, vol. 219, pp. 211–219, Jan. 2023, doi: 10.1016/j.procs.2023.01.283.

[10] X. Wang et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, May 2020, doi: 10.1016/j.csl.2020.101114.

[11] N. Tomashenko, Y. Khokhlov, and Y. Esteve, "Exploring Gaussian mixture model framework for speaker adaptation of deep neural network acoustic models," arXiv.org, Mar. 15, 2020. https://arxiv.org/abs/2003.06894.

[12] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, Feb. 2017, doi: 10.1016/j.csl.2017.01.001.

[13] L. Zhang and J. Yang, "A continuous liveness detection for voice authentication on smart devices," arXiv.org, Jun. 01, 2021. https://arxiv.org/abs/2106.00859

[14] X. Zhang, D. Cheng, P. Jia, Y. Dai and X. Xu, "An efficient android-based multimodal biometric authentication system with face and voice," *IEEE Access*, vol. 8, pp. 102757-102772, 2020, doi: 10.1109/ACCESS.2020.2999115.

[15] X. Wang, Z. Yan, R. Zhang, and P. Zhang, "Attacks and defenses in user authentication systems: A survey," *Journal of Network and Computer Applications*, vol. 188, p. 103080, May 2021, doi: 10.1016/j.jnca.2021.103080.