Deepfake Audio Detection in Voice Authentication: A Spectral and CNN-Based Comprehensive Review

Ali Osman Mohammed Salih
Department of Information Systems and Cyber Security, College of Computing and Information Technology, University of Bisha, Bisha 61922, P.O Box: 551, Saudi Arabia
aomohammed@ub.edu.sa (corresponding author)

Abdelmajid Hassan Mansour Emam
Department of Information Technology, College of Computing and Information Technology - Khulais, University of Jeddah, 2841 - Ad Duf, Khulays 25535 - 7419, Saudi Arabia
emam@uj.edu.sa

Alwalid Bashier Gism Elseed Ahmed
Department of Computer Science and Artificial Intelligence, College of Computing and Information Technology, University of Bisha, Bisha 61922, P.O Box: 551, Saudi Arabia
alwlidbasheer@ub.edu.sa

Mahmoud Khalifa
Department of Computer Science and Artificial Intelligence, College of Computing and Information Technology, University of Bisha, Bisha 61922, P.O Box: 551, Saudi Arabia
mkhalifa@ub.edu.sa

Abdelrazig Suliman
Department of Information Systems and Cyber Security, College of Computing and Information Technology, University of Bisha, Bisha 61922, P.O Box: 551, Saudi Arabia
alsaid@ub.edu.sa

Nissrein Babiker Mohammed Babiker
Department of Information Systems and Cyber Security, College of Computing and Information Technology, University of Bisha, Bisha 61922, P.O Box: 551, Saudi Arabia
nbbabaker@ub.edu.sa

Received: 14 July 2025 | Revised: 29 August 2025 | Accepted: 9 September 2025

 DOI: https://doi.org/10.48084/etasr.13400

ABSTRACT

As voice authentication systems become increasingly integral to critical domains such as banking, smart assistants, and remote identity verification, they face escalating threats from AI-generated audio, commonly referred to as deepfakes. These synthetic voices, produced through advanced text-to-speech and voice conversion technologies, can convincingly imitate human speech, thereby undermining the reliability and security of authentication frameworks. This study provides a comprehensive review of spectral-based techniques for deepfake audio detection, highlighting the roles of spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), and Constant-Q Transform (CQT) in exposing time-frequency anomalies. The integration of Convolutional Neural Network (CNN)-based spoof detection modules before identity verification is identified as a critical architectural strategy to enhance system resilience. This review also outlines the prevailing challenges, including vulnerability due to emerging generative models, limited interpretability of deep learning classifiers, and decreased robustness under realistic or noisy conditions.

To advance the field, this study emphasizes promising research directions such as hybrid modeling approaches, adversarial training techniques, and the development of multilingual open-access deepfake audio datasets. By critically synthesizing existing research, this review aims to inform the design of more robust, generalizable, and transparent voice authentication systems capable of surviving the evolving landscape of audio-based threats.

Keywords-audio deepfakes; voice authentication; spoof detection; spectral features; CNN; ASV spoof

I. INTRODUCTION

In recent years, voice authentication systems have become increasingly prevalent across diverse applications, ranging from banking and smart assistants to access control and remote verification. As these systems gain widespread adoption, the threats to their integrity have escalated accordingly. One of the most significant emerging challenges is AI-generated speech, commonly referred to as deepfake audio [1].

Voice authentication systems play a crucial role in ensuring secure access across various domains. However, the rise of deepfakes—synthetic audio generated through advanced Text-To-Speech (TTS) and voice conversion technologies—poses a serious threat to their effectiveness. These artificially synthesized voices can convincingly mimic human speakers, thereby compromising the reliability of traditional voice recognition mechanisms. The growing sophistication of these attacks increases the risk of exploitation, particularly in security-critical applications.

To mitigate these challenges, recent research has increasingly focused on spectral analysis techniques that leverage the unique time-frequency characteristics of speech. These methods allow for the detection of inconsistencies indicative of artificial manipulation [2, 3]. As illustrated in Figure 1, comparing spectrograms of real and deepfake audio reveals key differences in the time-frequency patterns exploited by spectral methods. The real audio (left) exhibits rich, slightly irregular harmonic structures over time, with dynamic energy variations appearing as vibrant yellow and green bands concentrated below 4000 Hz, reflecting natural vocal resonance. In contrast, the deepfake audio (right) displays smoother, overly uniform frequency bands with reduced intensity variation, primarily in the lower spectrum. The absence of high-frequency details and the unnatural consistency of the purple–blue regions serve as strong indicators of synthetic generation. These spectral cues form the foundation for the development of effective detection systems capable of distinguishing genuine from AI-generated speach.

This distinction between real and synthetic audio is fundamental to understanding the threat landscape. Table I provides a comparative overview of key characteristics differentiating real from AI-generated deepfake voice across multiple acoustic and perceptual dimensions.

Modern voice authentication systems incorporate spectral-based deepfake detection modules upstream of the identity verification process, thereby strengthening defenses against AI-generated spoofing attempts. Figure 2 presents a generalized architectural model of such systems, where the authentication pipeline is augmented with a deepfake detection stage before decision-making.

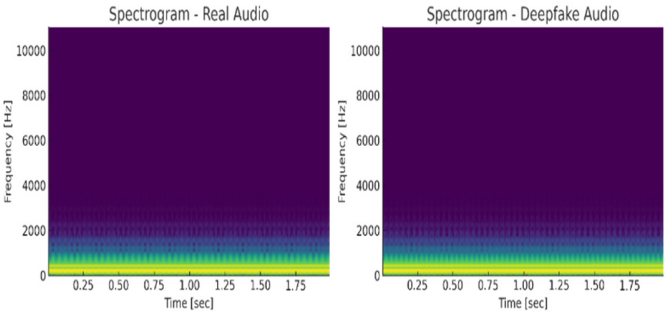

Fig. 1. Comparative Spectrograms of real (left) and deepfake (right) audio.

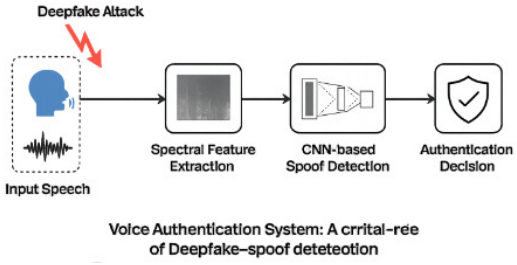

Fig. 2. A generalized architectural model of voice authentication systems integrating spectral-based spoof detection before the identity verification step.

TABLE I. COMPARISON BETWEEN REAL VOICE AND DEEPFAKE VOICE

Aspect	Real Voice	Deepfake Voice	References
Source	Recorded directly from a human speaker	Generated using AI techniques such as Text-to-Speech (TTS) or Voice Conversion	[4]
Spectral characteristics	Natural and complex variations in frequency and tone	Often exhibits uniform or unnatural spectral patterns due to its synthetic nature	[5]
Rhythm and intonation	Naturally variable, reflecting stress, emotion, and speaking context	It may sound overly consistent or lack expressive variation	[6]
Background noise	Typically includes ambient or environmental noise	Usually noise-free or contains artificial/repetitive background elements	[7]
Breathing and pauses	Contains natural breathing and irregular pauses	May lack breathing sounds or have artificially inserted pauses	[8]
Naturalness	Sounds authentic and is easily recognized as human speech	It may sound convincing, but it often lacks spontaneity or emotional depth	[9]

The process begins with an input speech signal, either genuine or synthetically generated through a deepfake attack. The signal first undergoes spectral feature extraction, where it is transformed into time–frequency representations such as spectrograms, MFCCs, or CQCCs. These spectral features highlight important speech characteristics and allow the system to capture subtle irregularities that may indicate tampering or synthesis. The extracted features are then passed to a CNN-based spoof detection module, which applies deep learning techniques to classify the speech as real or fake based on learned patterns. Only after successful spoof detection does the system proceed to the authentication decision stage to verify the speaker's identity. This layered approach significantly enhances the resilience of the system, filtering out manipulated audio before it can reach the core authentication engine.

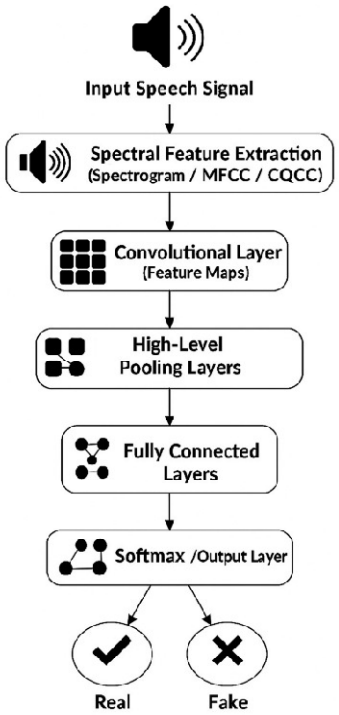

Fig. 3. Block diagram of a CNN-based spoof detection module.

Among the various spectral features employed, the Constant-Q Transform (CQT) has demonstrated notable effectiveness in anti-spoofing tasks [10]. CQT offers enhanced frequency resolution, particularly at lower frequencies—a region abundant with emotional and nuanced speech content, making it a valuable tool for distinguishing genuine from synthetic voices. Furthermore, combining multiple spectral features, including principal components and statistical descriptors, has been shown to improve detection performance against increasingly sophisticated spoofing attacks [11]. However, despite these advances, several challenges remain. Spectral-based detection methods are often sensitive to variations in recording conditions, background noise, and speaker diversity, which can significantly degrade their performance in real-world scenarios. Additionally, most models are trained on limited datasets and struggle to generalize to unseen generative models or languages, exposing a critical vulnerability in open-set conditions. This limitation raises concerns about their reliability in high-stakes real-world deployments.

Another notable challenge involves the interpretability of deep learning-based detectors, particularly CNNs. Although CNNs excel at extracting complex features from spectrograms, their decision-making processes often lack transparency. This opacity undermines user trust and complicates auditing in security-sensitive contexts. Moreover, the computational demands of processing high-resolution spectrograms can hinder practical deployment on edge or mobile devices.

Looking ahead, there is a clear need for detection frameworks that are not only robust and generalizable, but also interpretable. Future research directions include developing hybrid models that integrate spectral features with speaker embeddings and prosodic cues, as well as employing adversarial training strategies to enhance model resilience. Additionally, the creation of open-access, multilingual, and diverse deepfake datasets is essential for the thorough evaluation of system performance under evolving and realistic threat conditions.

In addition to CNN-based methods, traditional forensic techniques such as Linear Predictive Coding (LPC) can be incorporated for deepfake detection. LPC analyzes the spectral envelope of speech, capturing characteristic patterns that may differ between genuine and synthesized voices. Integrating LPC with modern deep learning approaches can enhance detection robustness, providing complementary insights for identifying manipulated audio [12].

The increasing sophistication and availability of audio deepfake tools have led to a notable increase in reported incidents of synthetic voice-based fraud. As shown in Figure 4, the estimated number of deepfake audio breaches has grown sharply between 2019 and 2024, reinforcing the urgency of developing more effective detection mechanisms [13]. Despite advances in detection, real-world deployment remains challenging due to variability in conditions and evolving attack methods. As observed in previous studies [14, 15], the number of reported audio deepfake incidents has shown a sharp increase. Figure 4 presents an estimate of this trend from 2019 to 2024, based on an aggregated interpretation of the discussed threat levels and case reports.

In summary, the architecture illustrated in this review presents a modern voice authentication pipeline in which incoming speech—whether genuine or synthetic—is processed through spectral feature extraction and CNN-based spoof detection before the final authentication decision. Spectral-based defenses serve as a pivotal component in reliably identifying manipulated or AI-generated inputs.

This review aimed to critically synthesize recent progress in spectral-based deepfake detection for voice authentication, highlighting key strengths, limitations, and research gaps. It also outlines future research directions needed to advance the field toward more trustworthy and deployable solutions.

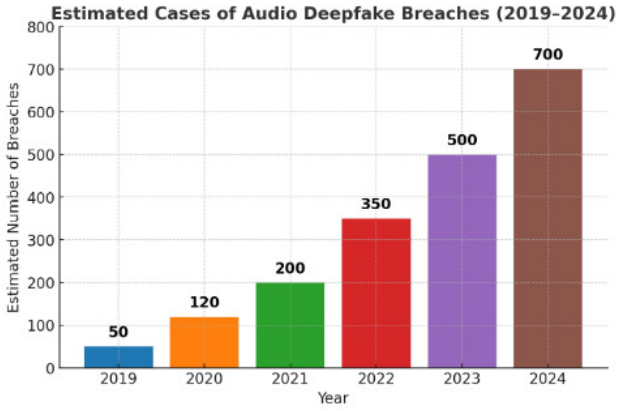

Fig. 4. An approximate estimation of the number of breaches involving deepfake audio attacks from 2019 to 2024, based on an analysis of data reported in [16, 17].

A. Motivation

As voice authentication becomes more prevalent in high-stakes applications such as banking and remote access control, audio deepfakes have emerged as a significant threat to biometric security. These AI-generated voices can closely mimic real speakers, enabling identity fraud, misinformation, and unauthorized access to systems [18, 19]. Despite the development of numerous detection approaches based on spectral analysis and deep learning [20], critical challenges persist—including poor generalization to unseen spoofing techniques [21], limited availability of diverse datasets [22], and lack of model interpretability [23]. These limitations highlight the urgent need for a critical review that synthesizes existing work, identifies knowledge gaps, and proposes new directions to enhance the robustness, transparency, and trustworthiness of voice authentication systems in the face of evolving deepfake threats.

B. Scientific Contribution

This review offers a structured and in-depth synthesis of recent advances in spectral-based audio deepfake detection, with a special emphasis on the use of CNNs in conjunction with spectral representations such as Mel spectrograms and Constant-Q Cepstral Coefficients (CQCC) [24]. Unlike previous review efforts [25], this study systematically classifies detection methods based on spectral features, model architectures, and spoofing attack types, providing a detailed comparative evaluation of their strengths and weaknesses. In particular, it addresses the often-overlooked dimension of model interpretability—a critical factor for building trust in AI systems—by advocating the integration of eXplainable AI (XAI) techniques into deepfake detection frameworks [26]. This study also identifies the main limitations in dataset diversity and language coverage, stressing the importance of multilingual, noisy, and real-world datasets to improve generalization and reliability [27]. Through these contributions, this review aims to guide both researchers and practitioners in designing transparent, adaptable, and resilient voice authentication systems that can withstand increasingly sophisticated audio deepfake attacks.

II. LITERATURE REVIEW

The rapid advancement of AI-driven voice synthesis has given rise to highly realistic audio deepfakes—synthetic voices capable of closely mimicking real human speech. These developments pose severe risks to voice authentication systems, privacy, and digital security. As a result, the detection of audio deepfakes has become a vital area of research, particularly in contexts requiring reliable speaker verification [28].

A. Spectral Features for Detection

Spectral features are essential for capturing the time–frequency characteristics of speech and revealing anomalies introduced during synthesis. The most prominent spectral representations used in deepfake detection include:

- Mel Spectrogram captures perceptually relevant resolution, enabling better detection of subtle artifacts in frequency bands aligned with the human auditory system. It is widely adopted due to its effectiveness in modeling speech energy patterns [29].
- Constant-Q Cepstral Coefficients (CQCC) provide high-resolution spectral analysis across multiple frequency bands, enhancing the detection of subtle inconsistencies in synthesized audio [30].
- Mel-Frequency Cepstral Coefficients (MFCC) are often used in combination with CQCC to improve robustness through feature fusion.

These features serve as input to deep learning models, which learn discriminative patterns for classifying speech as genuine or synthesized.

B. Deep Learning-Based Detection Approaches

CNNs have become the dominant architecture for detecting deepfake audio due to their ability to learn high-level spatial representations from spectrograms. Key approaches include:

- CNNs applied to Mel spectrograms or CQCCs: These models have demonstrated strong performance in distinguishing spoofed from genuine audio [31].
- End-to-end models: These architectures learn directly from raw audio waveforms, eliminating the need for handcrafted feature extraction.
- Interpretable CNN frameworks: Such models incorporate techniques like Layer-wise Relevance Propagation (LRP) to enhance transparency and support model auditing in high-stakes environments.

C. Current Challenges and Limitations

Despite notable progress, several challenges continue to hinder the real-world deployment of audio deepfake detection systems. First, many models exhibit poor generalization, failing to maintain performance when confronted with unseen spoofing techniques or datasets [32]. Second, the lack of interpretability remains a critical concern, as most deep learning-based detectors function as black boxes, limiting user trust and transparency [33]. Finally, limited dataset diversity constrains model effectiveness, as existing datasets often lack

sufficient variation in speakers, languages, and synthesis techniques, reducing their representativeness of real-world scenarios [34].

D. Future Research Directions

Addressing these challenges requires a shift toward more adaptable and interpretable detection systems. Promising directions include developing hybrid models that combine multiple spectral features while leveraging advanced architectures, such as transformers. Implementing XAI frameworks can improve trust and provide greater insight into model behavior. Additionally, expanding multilingual and real-world datasets is essential to enhance robustness and enable evaluation under diverse threat conditions. Building on these insights, this review provides a comparative analysis of state-of-the-art methods and offers guidance for the design of next-generation voice authentication systems.

III. METHODOLOGY

This review employed a structured and systematic method to investigate and synthesize recent advances in audio deepfake detection, with a particular focus on approaches that leverage spectral features and CNNs. The methodological framework was carefully designed to ensure objectivity, reproducibility, and comprehensive coverage of the relevant scientific literature. The process comprises five sequential stages, as described below.

A. Search Strategy and Data Sources

A well-defined and targeted search strategy was developed to identify high-quality peer-reviewed publications relevant to the study objectives. The search employed carefully selected combinations of keywords, including 'audio deepfake,' 'voice spoofing detection,' 'spectral features,' 'Mel spectrogram,' 'MFCC,' 'CQCC,' 'convolutional neural networks,' and 'ASVspoof dataset'. To ensure comprehensive coverage and scholarly credibility, literature searches were conducted across several well-established academic databases, namely:

- IEEE Xplore
- ScienceDirect
- SpringerLink
- ACM Digital Library
- Google Scholar

These databases were selected for their extensive coverage of publications in the fields of signal processing, artificial intelligence, machine learning, and cybersecurity.

B. Inclusion Criteria

To ensure methodological rigor and alignment with the review's research objectives, the following inclusion criteria were applied:

- Studies published between 2018 and 2024.
- Research explicitly focused on audio deepfake detection employing spectral features and deep learning models, with particular emphasis on CNNs.
- Use of benchmark datasets, including ASVspoof2019, ASVspoof2021, or Fake-or-Real.
- Reporting of quantitative evaluation metrics, such as Equal Error Rate (EER), accuracy, or Area Under the Curve (AUC).

C. Screening and Selection Procedure

The study selection process followed a structured, multi-stage pipeline to ensure both precision and relevance. Initially, studies were retrieved using keyword-driven searches. Duplicate records were removed, and the remaining titles and abstracts were screened for thematic alignment. Eligible studies then underwent a full-text review, during which critical methodological and technical attributes were extracted, including:

- Types of spectral features employed (e.g., Mel spectrogram, MFCC, CQCC).
- Model architectures implemented (e.g., CNN, ResNet, hybrid frameworks).
- Datasets utilized and evaluation protocols applied.
- Key contributions, performance metrics, and reported limitations [35].

To illustrate the common workflow identified across the reviewed studies, Figure 5 presents a generalized signal processing pipeline for spectral-based audio deepfake detection using CNNs. This pipeline includes stages from raw audio acquisition, preprocessing, and spectral feature extraction to CNN-based classification, ultimately producing a binary decision output (real or fake).

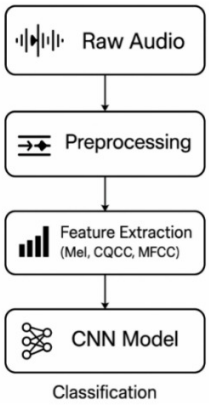

Fig. 5. Typical pipeline for spectral-based feature extraction and CNN-driven classification in audio deepfake detection systems.

D. Thematic Classification and Comparative Evaluation

The finalized set of studies was thematically categorized along three primary dimensions:

- Type of spectral representation used (e.g., MFCC, CQCC, Mel spectrogram).
- Architecture and complexity of the deep learning models implemented.

- Targeted spoofing scenarios, including logical access and physical access attacks.

This classification facilitated a comprehensive comparative analysis, yielding several key insights:

- Relative performance of spectral features, with Mel spectrograms frequently outperforming CQCC and MFCC across multiple benchmarks.
- Performance trends in model architectures, indicating that deeper CNN variants often achieve higher accuracy but exhibit greater susceptibility to overfitting.
- Persistent limitations, such as poor generalization to unseen spoofing techniques and strong dependence on specific datasets.
- Critical research gaps, including limited adoption of XAI techniques, insufficient evaluation of adversarial robustness, and a lack of diverse, multilingual, and real-world datasets [36].

E. Review Synthesis and Scholarly Contribution

By employing this systematic review method, this study provides a rigorous and integrative synthesis of advances in audio deepfake detection. The findings consolidate fragmented research, expose limitations within current detection paradigms, and outline strategic directions for future work. Ultimately, this review contributes to strengthening voice authentication security by providing a structured roadmap for developing more robust, generalizable, and interpretable deepfake detection systems.

F. Comparative Summary of Key Studies on Audio Deepfake Detection

1) Comparative Critical Analysis of Model Performance

Table II presents a comparative in-depth analysis of state-of-the-art deep learning architectures used for spoofing detection in Automatic Speaker Verification (ASV) systems. This table outlines the datasets utilized (D1-D5 correspond to successive ASVspoof challenge corpora), the spectral features adopted, the model paradigms (discriminative or generative), architectural configurations, and the principal performance indices—EER and tandem Detection Cost Function (t-DCF). Lower values of these metrics denote superior discriminative capacity in differentiating authentic from spoofed speech. Despite substantial advances facilitated by deep learning, the existing body of research exhibits notable constraints, including its predominant focus on a narrow set of languages and benchmark datasets. This limitation underscores an urgent need for the development of models with enhanced cross-lingual generalization, increased resilience under real-world acoustic conditions, and scalable, resource-efficient architectures capable of accommodating a broader spectrum of spoofing techniques.

2) Comparative Insights and Strategic Research Outlook

The ASVspoof2019 dataset comprises two primary protocols: Logical Access (LA) and Physical Access (PA). The LA protocol targets digitally generated spoofing attacks using Text-to-Speech (TTS) and Voice Conversion (VC) techniques, absent of any physical recording channel. In contrast, the PA protocol replicates real-world conditions by replaying audio through loudspeakers and microphones across diverse acoustic environments. This dual-protocol structure enables a comprehensive evaluation of spoof detection systems against both synthetic and replay-based attacks, establishing a robust benchmark to evaluate the performance of CNN-based detection algorithms.

The findings, summarized in Table IV, reveal a wide spectrum of deep learning strategies implemented for spoofing detection in ASV systems. Discriminative models—notably CNN, ResNet, and hybrid CNN-GRU architectures—consistently outperform classical and generative counterparts (e.g., Gaussian Mixture Models–GMM and Probabilistic Linear Discriminant Analysis–PLDA) across multiple benchmark datasets. This performance advantage is particularly pronounced when these models are paired with high-resolution spectral features such as CQCC, Mel-spectrogram, and CQT.

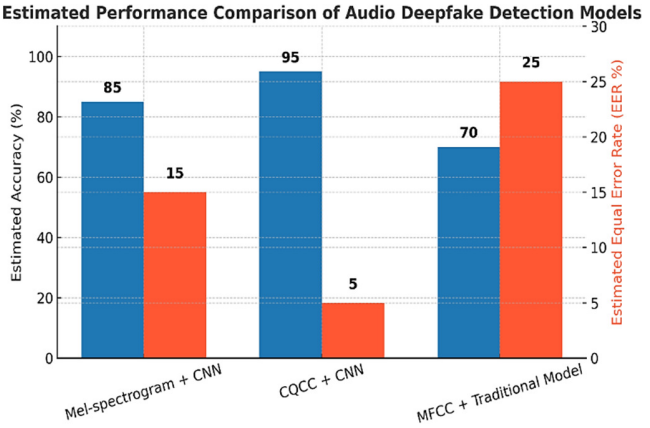

Fig. 6. Performance comparison of spectral-based models.

Performance metrics, including EER and t-DCF, demonstrate marked improvements in spoofing detection accuracy. For example, a ResNet architecture combined with CQCC achieved an EER of 0.59% on the ASVspoof2019 dataset (D5), indicating substantial robustness against advanced spoofing attacks. Still, persistent challenges remain, with the most important being limited cross-domain generalization to novel spoofing techniques and a lack of model interpretability, as most state-of-the-art approaches continue to operate as opaque black boxes. To overcome these challenges and enable the practical deployment of secure voice authentication systems, the following directions are recommended:

- Advancing spectral feature representations: Develop richer and more interpretable spectral features through deep learning methods that transcend traditional handcrafted approaches.
- Implementing hybrid modeling paradigms: Combine generative models (e.g., Variational Autoencoders–VAEs, Gaussian Mixture Models–GMMs) with discriminative architectures (e.g., ResNet, CNN+GRU) to enhance cross-domain generalization [42].

- Leveraging continual and self-supervised learning: Facilitate adaptive learning to address data drift and evolving spoofing techniques, reducing the need for frequent manual retraining.
- Enhancing transparency with XAI: Integrate techniques such as attention mechanisms and saliency maps to improve interpretability and strengthen user trust.
- Expanding and diversifying datasets: Incorporate multilingual, cross-channel, and noisy real-world recordings to improve model robustness and operational relevance.
- Employing comprehensive evaluation metrics: Utilize t-DCF in conjunction with EER to capture a more nuanced understanding of system performance in practical deployment scenarios.

IV. DISCUSSION

This review investigated state-of-the-art audio deepfake detection methodologies, emphasizing spectral feature representations—Mel-spectrograms, CQCC, and MFCC—and their integration with deep learning architectures, particularly CNNs, using benchmark datasets such as ASVspoof2019.

Mel-spectrograms capture both spectral and temporal cues that align with human auditory perception, facilitating CNNs in detecting subtle inconsistencies characteristic of synthetic speech. CQCC provides enhanced frequency resolution and increased sensitivity to fine-grained distortions introduced by advanced voice synthesis techniques, thereby bolstering robustness against sophisticated spoofing attacks. MFCC, while offering a balanced spectral representation, exhibits comparatively lower standalone performance, likely due to limited sensitivity to emotional and prosodic cues, which are critical for identifying subtle deepfake artifacts [43-45]. Despite progress, some challenges remain:

- Limited generalization to unseen spoofing methods and out-of-distribution audio.
- CNNs' black-box nature limits interpretability, crucial in security applications.
- High computational costs hinder real-time deployment.

Future research should focus on hybrid models incorporating prosodic and speaker embeddings, XAI tools, and expanding large-scale multilingual, noise-rich datasets to better represent real-world conditions.

TABLE II. SUMMARY OF DL MODELS FOR ASV SPOOFING DETECTION

Study	Dataset	Features	Model type	Architecture	Modeling approach	Metric (EER / t-DCF)
[37]	D1	raw-waveform	end-to-end	CNN	Disc	EER = 0.157
[38]	D3.1	embedding	classification	GMM	Gen	EER = 6.4
[39]	D5	embedding	classification	PLDA	Gen	EER = 2.23, t-DCF = 0.0614
[40]	D4, D5	CQCC, spectrogram	classification	ResNet, SeNet	Disc	EER = 0.59, t-DCF = 0.016
[41]	D5	spectrogram	classification	CNN+GRU	Disc	EER = 2.45, t-DCF = 0.0570

TABLE III. COMPARATIVE SUMMARY OF SPECTRAL FEATURES IN AUDIO DEEPFAKE DETECTION

Feature type	Time resolution	Frequency resolution	Robustness to noise	Commonly used with
Mel-spectrogram	High	Moderate	Moderate	CNN, RNN
CQCC	Variable	High	High	GMM, CNN
MFCC	Moderate	Moderate	Moderate	SVM, HMM

Deep learning models, particularly CNNs, consistently outperform classical models such as SVMs and HMMs

TABLE IV. QUALITATIVE PERFORMANCE COMPARISON OF FEATURE–MODEL COMBINATIONS

Model Combination	Expected Accuracy	Relative EER (Lower is Better)	Reference
Mel-spectrogram + CNN	High	Moderate	[46]
CQCC + CNN	Very High	Low	[47]
MFCC + Traditional model	Moderate	High	[48]

Hybrid approaches combining multiple spectral features have demonstrated enhanced detection across diverse spoofing types

V. CONCLUSION

This review provides a comprehensive analysis of recent advances in spectral-based audio deepfake detection, particularly within voice authentication systems. By examining spectral features such as Mel-spectrograms, MFCC, and CQCC, in combination with deep learning architectures such as CNNs and hybrid models, it is evident that spectral cues remain fundamental for detecting subtle artifacts introduced by AI-generated speech. Comparative evaluations demonstrate that models integrating CQCC with advanced CNN architectures, such as ResNet, achieve superior detection performance, particularly in controlled benchmark scenarios such as ASVspoof 2019.

However, this review also highlights systemic challenges in current research, including limited robustness under real-world conditions, poor generalization to novel or cross-domain spoofing attacks, and insufficient attention to model interpretability and explainability. Moreover, the scarcity of multilingual, noisy, and publicly available datasets continues to hinder practical progress. Moving forward, the field requires

hybrid detection frameworks to fuse spectral, prosodic, and embedding-level representations, incorporate XAI techniques to enhance trust and transparency, and establish standardized evaluation protocols alongside open-access corpora to facilitate reproducibility and cross-system benchmarking. Addressing these gaps will pave the way for resilient, generalizable, and interpretable voice authentication systems capable of effectively countering evolving deepfake audio threats.

ACKNOWLEDGMENT

The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

REFERENCES

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, Feb. 2015, https://doi.org/10.1016/j.specom.2014.10.005.

[2] H. Shi, X. Shi, S. Dogan, S. Alzubi, T. Huang, and Y. Zhang, "Benchmarking Audio Deepfake Detection Robustness in Real-world Communication Scenarios." arXiv, 2025, https://doi.org/10.48550/ARXIV.2504.12423.

[3] T. Kinnunen *et al.*, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Interspeech 2017*, Aug. 2017, pp. 2–6, https://doi.org/10.21437/Interspeech.2017-1111.

[4] S. Barrington, R. Barua, G. Koorma, and H. Farid, "Single and Multi-Speaker Cloned Voice Detection: From Perceptual to Learned Features," in *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, Nürnberg, Germany, Dec. 2023, pp. 1–6, https://doi.org/10.1109/WIFS58808.2023.10374911.

[5] A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva, "Training-Free Deepfake Voice Recognition by Leveraging Large-Scale Pre-Trained Models," in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, Baiona, Spain, Jun. 2024, pp. 289–294, https://doi.org/10.1145/3658664.3659662.

[6] J. Yi *et al.*, "SceneFake: An initial dataset and benchmarks for scene fake audio detection," *Pattern Recognition*, vol. 152, Aug. 2024, Art. no. 110468, https://doi.org/10.1016/j.patcog.2024.110468.

[7] S. Sarkar, A. Gupta, A. Ghosh, and S. Ganesan, "DeepFake Classification Using Fine-Tuned Wave2Vec2.0," in *Artificial Intelligence and Speech Technology*, vol. 2390, A. Sharma and R. Rani, Eds. Springer Nature Switzerland, 2025, pp. 67–78.

[8] X. Zhang, J. Yi, C. Wang, C. Y. Zhang, S. Zeng, and J. Tao, "What to Remember: Self-Adaptive Continual Learning for Audio Deepfake Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 19569–19577, Mar. 2024, https://doi.org/10.1609/aaai.v38i17.29929.

[9] P. Balasubramanian *et al.*, "Generative AI for cyber threat intelligence: applications, challenges, and analysis of real-world case studies," *Artificial Intelligence Review*, vol. 58, no. 11, Aug. 2025, Art. no. 336, https://doi.org/10.1007/s10462-025-11338-z.

[10] A. Nautsch *et al.*, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, Apr. 2021, https://doi.org/10.1109/TBIOM.2021.3059479.

[11] L. M. D. Souza, R. C. Guido, R. C. Contreras, M. S. Viana, and M. A. D. S. Bongarti, "Improving Voice Spoofing Detection Through Extensive Analysis of Multicepstral Feature Reduction," *Sensors*, vol. 25, no. 15, Aug. 2025, Art. no. 4821, https://doi.org/10.3390/s25154821.

[12] O. A. Shaaban and R. Yildirim, "Audio Deepfake Detection Using Deep Learning," *Engineering Reports*, vol. 7, no. 3, Mar. 2025, Art. no. e70087, https://doi.org/10.1002/eng2.70087.

[13] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, no. 7, Mar. 2025, Art. no. 1989, https://doi.org/10.3390/s25071989.

[14] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: generation, detection, and applications," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 3, pp. 219–289, Sep. 2022, https://doi.org/10.1007/s13735-022-00241-w.

[15] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, Aug. 2020, https://doi.org/10.1109/JSTSP.2020.3002101.

[16] A. Naitali, M. Ridouani, F. Salahdine, and N. Kaabouch, "Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions," *Computers*, vol. 12, no. 10, Oct. 2023, Art. no. 216, https://doi.org/10.3390/computers12100216.

[17] A. Firc, K. Malinka, and P. Hanáček, "Evaluation framework for deepfake speech detection: a comparative study of state-of-the-art deepfake speech detectors," *Cybersecurity*, vol. 8, no. 1, Aug. 2025, Art. no. 50, https://doi.org/10.1186/s42400-024-00346-1.

[18] J. Bateman, Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios. Carnegie - Endowment for International Peace, 2020.

[19] T. Hidar, A. A. E. Kalam, and Y. Lamtigui, "Securing Biometric Authentication Systems: A Hybrid Methodology for DeepFake Detection and Response," in *Proceedings of the 4th International Conference on Advances in Communication Technology and Computer Engineering (ICACTCE'24)*, vol. 1312, 2025, pp. 369–380.

[20] A. Raza, K. Munir, and M. Almutairi, "A Novel Deep Learning Approach for Deepfake Image Detection," *Applied Sciences*, vol. 12, no. 19, Sep. 2022, Art. no. 9820, https://doi.org/10.3390/app12199820.

[21] Z. Wu *et al.*, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, Jun. 2017, https://doi.org/10.1109/JSTSP.2017.2671435.

[22] G. Ali, J. Rashid, M. Rameez Ul Hussnain, M. Usman Tariq, A. Ghani, and D. Kwak, "Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection," *IEEE Access*, vol. 12, pp. 149940–149959, 2024, https://doi.org/10.1109/ACCESS.2024.3457866.

[23] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023, https://doi.org/10.1007/s10489-022-03766-z.

[24] J. L. Narayana Budati, A. Bharathi Jadam, and R. Malleboyina, "Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics," in *2025 6th International Conference for Emerging Technology (INCET)*, Belgaum, India, May 2025, pp. 1–7, https://doi.org/10.1109/INCET64471.2025.11140952.

[25] N. Mansoor and A. I. Iliev, "Explainable AI for DeepFake Detection," *Applied Sciences*, vol. 15, no. 2, Jan. 2025, Art. no. 725, https://doi.org/10.3390/app15020725.

[26] W. A. Jbara, N. A. H. K. Hussein, and J. H. Soud, "Deepfake Detection in Video and Audio Clips: A Comprehensive Survey and Analysis," *Mesopotamian Journal of CyberSecurity*, vol. 4, no. 3, pp. 233–250, Dec. 2024, https://doi.org/10.58496/MJCS/2024/025.

[27] R. Rini, "Deepfakes and the Epistemic Backstop," *Philosophers' Imprint*, vol. 20, no. 24, pp. 1–16, 2020.

[28] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Interspeech 2017*, Aug. 2017, pp. 3642–3646, https://doi.org/10.21437/Interspeech.2017-1428.

[29] R. K. Bhukya, A. Raj, and A. Kumar, "ASVSpoof 2021: Detecting Spoofed Utterances Through Hybrid Features," *APSIPA Transactions on Signal and Information Processing*, vol. 14, no. 1, 2025, https://doi.org/10.1561/116.20250026.

[30] E. B. Da Conceicao Mahangue, A. K. Sharma, and K. Gupta, "Systematic Review on Detection of Deepfake Attack using Machine Learning," in *2025 12th International Conference on Computing for Sustainable Global Development (INDIACom)*, Delhi, India, Apr. 2025, pp. 1–5, https://doi.org/10.23919/INDIACom66777.2025.11115308.

[31] B. Chettri, E. Benetos, and B. L. T. Sturm, "Dataset Artefacts in Anti-Spoofing Systems: A Case Study on the ASVspoof 2017 Benchmark," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3018–3028, 2020, https://doi.org/10.1109/TASLP.2020.3036777.

[32] M. Li, Y. Ahmadiadli, and X. P. Zhang, "A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, Lisboa, Portugal, Oct. 2022, pp. 35–41, https://doi.org/10.1145/3552466.3556523.

[33] E. Şahin, N. N. Arslan, and D. Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Computing and Applications*, vol. 37, no. 2, pp. 859–965, Jan. 2025, https://doi.org/10.1007/s00521-024-10437-2.

[34] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "SpeechFake: A Large-Scale Multilingual Speech Deepfake Dataset Incorporating Cutting-Edge Generation Methods," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, 2025, pp. 9985–9998, https://doi.org/10.18653/v1/2025.acl-long.493.

[35] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "End-to-End convolutional neural network-based voice presentation attack detection," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, USA, Oct. 2017, pp. 335–341, https://doi.org/10.1109/BTAS.2017.8272715.

[36] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric," in *Interspeech 2018*, Sep. 2018, pp. 671–675, https://doi.org/10.21437/Interspeech.2018-1819.

[37] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection," in *Interspeech 2019*, Sep. 2019, pp. 1068–1072, https://doi.org/10.21437/Interspeech.2019-2212.

[38] C. I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual neTworks." arXiv, 2019, https://doi.org/10.48550/ARXIV.1904.01120.

[39] J. Jung, H. Shim, H. S. Heo, and H. J. Yu, "Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 Challenge." arXiv, Jul. 17, 2019, https://doi.org/10.48550/arXiv.1904.10134.

[40] J. Zhan, Z. Pu, W. Jiang, J. Wu, and Y. Yang, "Detecting Spoofed Speeches via Segment-Based Word CQCC and Average ZCR for Embedded Systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 11, pp. 3862–3873, Nov. 2022, https://doi.org/10.1109/TCAD.2022.3197531.

[41] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, Jan. 2019, https://doi.org/10.22215/timreview/1282.

[42] T. Kirchengast, "Deepfakes and image manipulation: criminalisation and control," *Information & Communications Technology Law*, vol. 29, no. 3, pp. 308–323, Sep. 2020, https://doi.org/10.1080/13600834.2020.1794615.

[43] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio Replay Attack Detection with Deep Learning Frameworks," in *Interspeech 2017*, Aug. 2017, pp. 82–86, https://doi.org/10.21437/Interspeech.2017-360.

[44] T. Kinnunen *et al.*, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification." arXiv, 2018, https://doi.org/10.48550/ARXIV.1804.09618.

[45] K. Borodin, V. Kudryavtsev, G. Mkrtchian, and M. Gorodnichev, "Capsule-based and TCN-based Approaches for Spoofing Detection in Voice Biometry," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18409–18414, Dec. 2024, https://doi.org/10.48084/etasr.8906.

[46] M. K. Z. Bajwa, A. Castiglione, and C. Pero, "Mel Spectrogram-Based CNN Framework for Explainable Audio Deepfake Detection," in *Advanced Information Networking and Applications*, vol. 252, L. Barolli, Ed. Springer Nature Switzerland, 2025, pp. 407–416.

[47] T. M. Wani, S. A. A. Qadri, D. Comminiello, and I. Amerini, "Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation," in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, Baiona, Spain, Jun. 2024, pp. 271–276, https://doi.org/10.1145/3658664.3659647.

[48] A. Hamza *et al.*, "Deepfake Audio Detection via MFCC Features Using Machine Learning," *IEEE Access*, vol. 10, pp. 134018–134028, 2022, https://doi.org/10.1109/ACCESS.2022.3231480.