*International Journal of Basic and Applied Sciences*, 14 (5) (2025) 20-25

International Journal of Basic and Applied Sciences

*Website: www.sciencepubco.com/index.php/IJBAS*
*https://doi.org/10.14419/ct23r415*
***Research paper***

# Comparative Analysis Of ML/DL Models for Voice Spoofing Detection

**Rekha Rani [1] *, Bal Kishan [2]**

[1] *Research Scholar, Department of Computer Science, Maharshi Dayanand University, Rohtak, Haryana, India*
[2] *Associate Professor, Department of Computer Science, Maharshi Dayanand University, Rohtak, Haryana*
**Corresponding author E-mail: singhrekha964@gmail.com*

*Received: July 13, 2025, Accepted: August 19, 2025, Published: September 1, 2025*

**Abstract**

Spoofing of voice is an alarming threat to voice-security systems, especially in high-stakes areas like banking, smart home gadgets, customer support, and virtual assistants. In this paper, the authors submit an inclusive comparative review of spoof detection methods with a description of how they evolved from classical machine learning to the latest deep learning and transformer-based models. The performance of different models is measured across benchmark datasets with common metrics like Equal Error Rate (EER) and tandem Detection Cost Function (t-DCF). Interestingly, the increase in the effectiveness of Transformer architectures for spoofed audio detection is emphasized. The ultimate goal of the present research is to assist both professionals and academics in selecting and developing dependable and safe voice authentication systems.

***Keywords***: *Spoofing Attacks; Machine Learning; Deep Learning.*

## 1. Introduction

Technology used for speech recognition has come a long way in the last few years and is now being used in several useful digital systems. This technology can be used for many things, like smart virtual helpers, automatic customer service, and payments. Because these systems are used so much, security holes like voice faking attacks become easier to find and exploit. Spoofing is when someone changes or imitates someone else's voice to trick voice verification systems. In the past few years, voice recognition technology has gotten better and is now an important part of most computer systems. This technology is used in many areas, like banking, customer service, and smart helpers. These systems are more likely to be attacked by hackers because they are used so often, especially voice faking attacks. Spoofing is when someone changes or imitates the voice of another person to trick a voice authentication system. Voice conversion, which records words and changes them to sound like someone else's voice, replay attacks, and synthetic speech, which uses text-to-speech technologies, are all examples of these kinds of attacks. These kinds of problems are looked at to show how deep learning and machine learning models can be used to make safe systems.

Initially, the experiment employed typical machine learning techniques, which employed the Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) with hand-crafted features like Constant-Q Cepstral Coefficient (CQCC) and Mel-Frequency Cepstral Coefficients (MFCC). [Wu et al. (2015), Hasan et al. (2017)]. These methods worked well for querying simple signal patterns but generally suffered from noise, many languages, or different datasets. Handcrafted features performed worse.

Deep learning revolutionized the discipline with models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) [Zhang et al. (2020)], or Long Short-Term Memory (LSTM) [Li et al. (2021)]. These algorithms enhanced accuracy and efficiency by extracting complex features from raw audio data. Recent work has verified that using Transformer-based models, which can extract key features from audio automatically without any pre-processing, enhances task performance and generalization [Chen et al., 2023]. Apart from the popular ASVspoof 2019 corpus, VoxCeleb, Voice Spoofing Detection Corpus (VSDC), and Voice Conversion Challenge (VCC) present other test factors.

Even with all this improvement, it's still hard to tell whether someone is trying to fake their speech, especially when there are several languages involved, not many resources, and the need for real-time execution. Moreover, preserving user anonymity, reducing evaluation costs, and protecting model generalizability from emerging attack vectors continue to be significant challenges. Zhao et al. conducted a study that provides a comparative review of various machine learning (ML) methodologies for speech spoofing detection, emphasizing the identification of pragmatic trade-offs and the direction of future research towards the development of reliable and efficient autonomous systems [Zhao et al., 2022]. Recent research also highlights privacy-safeguarding methods like federated learning and secure model updates [Zhao et al., 2022], which provide collaborative training without endangering sensitive user information.

This work provides a comparative survey of various machine learning solutions towards voice spoofing detection, evaluated on multiple benchmark datasets (ASVspoof 2019, VoxCeleb, VSDC, VCC) using metrics such as EER, t-DCF, and computational cost. Here, we outline the benefits and drawbacks of each method, emphasizing trade-offs between accuracy and efficiency, while highlighting the novelty of our review in combining performance, robustness, and computational analysis in a single comparative framework. The growing influence of Transformer-based architectures is also discussed in the context of enhancing voice security.
Below Fig. 1, illustrates the overview of evolution from the traditional Machine Learning (handcrafted + classifier) approach to modern Deep Learning approaches (CNNs, LSTMs, and Transformers), just by learning directly from raw audio features.

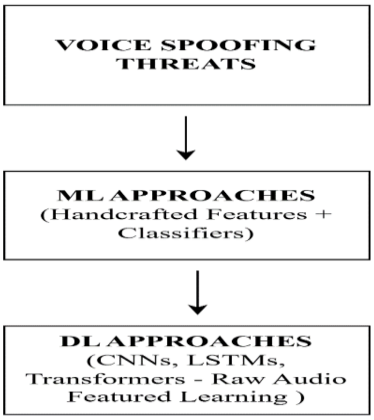


**Fig. 1:** Overview of the Evolution from Traditional ML to Modern DL Approaches in Voice Spoofing Detection.

## 2. Background and related work

Voice spoofing detection has improved significantly in the last ten years. Initially, methods used SVMs or GMMs to recognize handcrafted features like Mel-Frequency Cepstral Coefficients (MFCC), Constant-Q Cepstral Coefficients (CQCC), and Linear Predictive Coding Coefficients (LPCC). For instance, Wu et al. (2015) used a GMM classifier with CQCC features to recognize replay attacks on the ASVspoof 2015 dataset, attaining good performance on known attacks but decreased robustness against novel ones.
In this way, Hasan et al. (2013) combine SVMs and MFCCs to show great results in a lab. In general, such fixed features are not changed by noise or under-the-table spoofing patterns, state Kinnunen et al. (2017). These kinds of methods limit flexibility since they often fail when there is real-time noise, several languages, or cross-data set trials, even if they are easier to understand and less expensive to compute.
Benchmark corpora are very important for setting up and testing ways to identify people. The ASVspoof Challenges (2015–2021), with standardized datasets, including voice conversion, replay, and synthetic speech for a range of attacking strategies, were the most often deployed in practice. To make it easier to compare studies, datasets like WaveFake, VoxCeleb, VSDC (Voice Spoofing Detection Corpus), and VCC (Voice Conversion Challenge) have added multilingual fluctuation, noisy pathways, and patchy spoof sample variation as test features. This means that the testing goes beyond ASVspoof.
Assessment of these systems has conventionally used measures like Equal Error Rate (EER) and tandem Detection Cost Function (t-DCF). EER yields a scalar value such that the false rejection rate equals the false acceptance rate, and model comparison is straightforward, but it does not reflect performance differences across various spoof types or varying operating conditions. The t-DCF measure, however, includes both spoofing detection and ASV mistakes, and as such is more applicable to integrated systems, although more difficult to interpret. Although accuracy, precision, recall, and F1-score are sometimes printed, these are less standardized in this field. Recent work also suggests adding computational complexity, memory usage, and inference latency as additional metrics to complement the evaluation of real-time deployability, particularly for IoT and mobile platforms (Zhao et al., 2022).
To mitigate these constraints, deep learning (DL) has become a prevalent method for voice spoofing detection. Architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) have proven to detect better, especially for synthetic and novel attacks, by learning automatically hierarchical and temporal dependencies from raw waveforms or spectrogram inputs. Nonetheless, CNNs can perform poorly when fine-grained spectral cues are masked by background noise, whereas LSTMs can overfit in low-resource or multilingual scenarios (Jung et al., 2022). Both architectures often necessitate large-scale annotated datasets, e.g., ASVspoof 2017 and ASVspoof 2019, and high computational resources both for training and inference, hindering real-time or on-device deployment. Lightweight transformations, such as MobileNet variants, model pruning, and quantization, have been investigated to reduce latency and memory consumption. These methods try to preserve most of the detection accuracy while allowing deployment on edge devices with constrained resources. Such efficiency improvements are now regarded as a necessary condition for latency-sensitive applications.
The most recent innovations in voice signal analysis are transformer-based models, including Wav2Vec 2.0 and RawNet2 [Baevski et al. (2020); Tak et al. (2021)]. These models benefit from self-attention and self-supervised pre-training on vast amounts of unlabelled speech signals, allowing them to extract both short-term acoustic information and long-term dependency essential for spoof detection. When tested on benchmarking datasets such as ASVspoof 2019, these models perform outstandingly, picking up even slight voice variations and sophisticated spoof attacks. Chen et al. (2023) tried to fill a big gap in robustness that recent surveys found. They showed that Transformer frameworks operate well even in noisy and multilingual assessment environments, which is when CNN and GMM models have trouble. They can learn from a lot of speech that doesn't contain labels, but they are especially beneficial in multilingual contexts and when there aren't many resources. Some of the big computational resources needed to limit the deployment are long training times, high memory requirements, and slower processing rates. Some frequent improvements are privacy-protecting features like federated learning, which lets you train without giving away any personal information, and smaller Transformer models (like DistilWav2Vec), which save money on computing.
Nowadays, there is a latest trend of Lightweight Transformer models (like DistilWav2Vec) which simplify the processing costs and protect the framework's privacy, like federated learning, which promotes cooperative training without disclosing private audio first. Multimodal and hybrid spoofing detection is one technique that boosts resistance by pairing audio with an additional instrument, like lip recognition or fingerprinting. Adversarial training has additionally been utilized to increase susceptibility against perturbation-based attacks. Further,

explainability continues to improve; summarizing model results with techniques like SHAP and LIME may enhance the credibility of automated detection, which is necessary for high-stakes use cases like banking or legal forensics. Although ASVspoof remains the mainstay, most recent developments include an element of heterogeneity to the agenda: datasets like VoxCeleb and the Voice Spoofing Detection Challenge (VSDC) add diverse accents, noisy audio, and scarce spoof samples to the mix, adding a better simulation of stress testing for spoof detectors.
The following Table 1 summarizes different voice spoofing detection methods, showing how they perform on various datasets and metrics. This structured format helps identify not only the most accurate methods but also the ones that are practical under different deployment constraints.

**Table 1:** Comparative Overview of Voice Spoofing Detection Methods

| Years (citations) | Key Methods | Feature Extraction | Classifiers/ Models | Typical Datasets | Strength | Limitations | Performance Metrics |
|---|---|---|---|---|---|---|---|
| Pre-2015 [Wu et al. (2015), Hasan et al. (2013), Kinnunen et al (2017)] | Traditional ML | Handcrafted (MFCC, CQCC) | SVM, GMM | ASVspoof 2015 | Simple, understandable, low compute | Poor generalization, weak temporal modelling | EER, t-DCF. Computational cost |
| 2015-2020 [Various CNN, LSTM studies] | Deep Learning | Learned hierarchical (spectrogram, raw waveform ) | CNN, RNN, LSTM | ASVspoof 2017, 2019 | Automatic feature learning, improved temporal modelling | Data and compute-intensive, less robust in a noisy environment | EER, t-DCF, Inference time |
| 2020+ [Baevski et al. (2020), Jung et al. (2021)] | Transformer–based | Self-attention, raw waveform | Wav2Vec 2.0, RawNet2 | ASVspoof 2019, VoxCeleb. VSDC | Models long-range dependencies, resistant to novel attacks | Complex, resource-intensive, high GPU memory usage | EER, t-DCF, robustness score, latency |
| 2023+ (Emerging Trends) [Chen et al. (2023), Tak et al. (2023), Lavrentyeva et al. (2024)] | Hybrid & Multimodal / Advanced Trends | Multi-source (audio + visual), self-supervised embeddings, adversarial perturbations | Hybrid CNN-Transformer, SSL models (HuBERT, XLS-R), multimodal fusion networks [19] | ASVspoof 2021, VCC2020, Deepfake Detection Challenge, cross-dataset evaluation | Strengths of many techniques combined, cross-modal robustness, few-shot ability, privacy-preserving promise via federated learning | Complex deployment; synchronization problem, limited public multimodel spoof detection data | EER, t-DCF, robustness measures, resource consumption, cross-dataset accuracy |

## 3. Methodology

### 3.1. Review process

This study conducts a systematic review of voice spoofing detection methods ranging from conventional machine learning, deep learning to Transformer-based models. We aim to integrate results across different peer-reviewed research, highlighting main evaluation metrics and datasets, including ASVspoof 2019, VoxCeleb, and VSDC, to illustrate the evolution and comparative performance of spoofing detection systems across diverse conditions.
We followed a systematic approach as indicated in Fig. 2, which outlines the process from dataset selection and feature extraction to model categorization and performance evaluation. The review includes both accuracy-oriented and computationally focused comparisons, enabling a balanced perspective on real-world deployment feasibility.

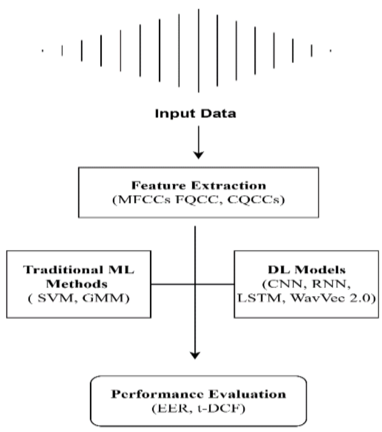


**Fig. 1:** Methodological Workflow for Voice Spoofing Detection.

The workflow begins with input audio data from multiple benchmark datasets, followed by feature extraction using MFCCs, FQCC, and CQCCs. Models are categorized into traditional ML methods (SVM, GMM) and deep learning approaches (CNN, RNN, LSTM, Transformer-based Wav2Vec 2.0), with performance assessed using Equal Error Rate (EER) and tandem Detection Cost Function (t-DCF).

### 3.2. Evaluation metrics

We rely on two common metrics popularly used in ASVspoof evaluations:

- Equal Error Rate (EER): The point at which false acceptance and false rejection are the same, lower the better. Lower values indicate better detection accuracy, and the metric is particularly useful for comparing systems across varied spoofing conditions.

- Tandem Detection Cost Function (t-DCF): By combining detection and verification errors into a single metric, it allows for the assessment of trade-offs between security and usability. Lower t-DCF values reflect better integration of spoof detection into automatic speaker verification (ASV) systems.

### 3.3. Comparative performance analysis

We provide EER and t-DCF values from well-known studies, illustrating sequential improvements in model performance. These result highlights the shift from handcrafted-feature ML systems to end-to-end DL and Transformer-based models, showing consistent gains in both accuracy and system robustness..

Transformer architectures such as Wav2Vec 2.0 demonstrate the lowest error rates, underscoring their potential for real-world deployment in challenging spoofing scenarios.

**Table 2:** Comparing Model Performance on ASVSPOOF 2019 Logical Access Dataset

| Citation’s | Model | EER (%) | t-DCF |
|---|---|---|---|
| Wu et al., 2015 [1] | GMM + CQCC | 9.87 | 0.48 |
| Hasan et al., 2013 [2] | SVM + MFCC | 7.21 | N/A |
| Tak et al., 2021 [4] | CNN (RawNet2) | 2.19 | 0.11 |
| Zhang et al., 2020 [7] | RNN | 3.85 | 0.21 |
| Li et al., 2021 [8] | LSTM | 1.63 | 0.09 |
| Jung et al., 2022 [9] | Wav2Vec 2.0 | 0.12 | 0.05 |

### 3.4. Visual insights

The dramatic drop in EER and t-DCF over time is shown in Figs 3 and 4, with Transformer-based and self-supervised models, showing the steepest performance gains.

The performance trends are visually represented in Figures 3 and 4, showing a sharp decline in both EER and t-DCF as models progress from traditional ML to advanced DL and Transformer-based approaches.

Figure 3 presents EER comparisons across spoofing detection models on the ASVspoof 2019 Logical Access dataset. The steep downward trend demonstrates the significant efficiency gains achieved through deep learning and Transformer-based architectures over earlier hand-crafted-feature methods.

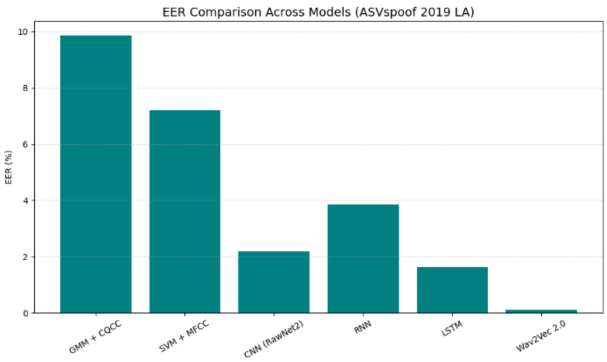


**Fig. 3:** Comparing EERs in Different Models.

Figure 4 depicts t-DCF comparisons across the same set of models. Wav2Vec 2.0 achieves the lowest detection cost, reflecting its superior balance between spoof detection accuracy and speaker verification reliability.

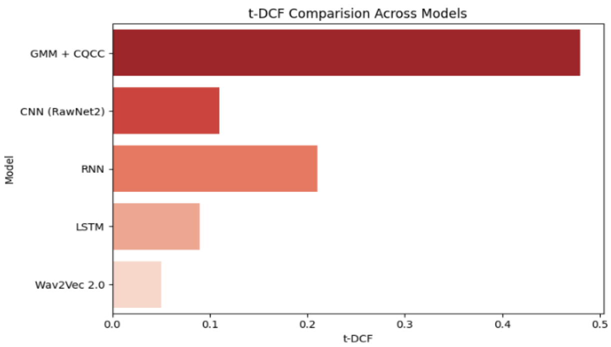


**Fig. 2:** Comparing t-DCF between Models.

## 4. Results and discussion

An analysis of different approaches to spoofing detection demonstrates a progressive improvement in model performance evaluated in parallel with advancements in model architecture. These improvements are consistently validated through standardized evaluation metrics such as Equal Error Rate (EER) and tandem Detection Cost Function (t-DCF), enabling fair comparisons across studies.

Classical machine learning approaches based on hand-crafted feature extraction (e.g., MFCC, CQCC) and a chosen classifier (e.g., SVM, GMM) offer reasonable baseline detection against attacks but struggle with more complex spoofing attacks and varying audio settings.

Their limitations are particularly evident in noisy environments, cross-channel settings, and multilingual contexts — scenarios common in real-world deployments but underrepresented in curated benchmark datasets.
Detection accuracy is significantly enhanced using deep learning models, including convolutional and recurrent neural networks (CNNs and LSTMs), as they automatically extract spatial and temporal features from the data, with recent multi-scale and cross-layer fusion strategies further improving feature richness [16]. But again, as with earlier models, these too struggle to accommodate fresh or more advanced spoofing attacks. But they need more computational and memory resources, which can cause latency when run on resource-limited or edge devices — mapping a long-standing trade-off between performance and operational efficiency. And again, as with their ML counterparts, DL models also struggle to learn to accommodate entirely new or more advanced spoofing strategies.
Transformer-based models are the latest development, utilizing self-attention mechanisms and self-supervised learning, to capture long-range dependencies and minute speech patterns. Architectures consistently achieve the lowest error rates and detection costs. This literature review has shown that the transformer models (such as Wav2Vec 2.0 and HuBERT), which are the most recent models based on self-attention, consistently achieve the lowest error rates and detection costs, demonstrating superior cross-dataset generalization even under mismatched training and testing conditions, when compared to traditional machine learning (ML) and deep learning (DL) models along with their classical counterparts. This robustness is critical for real-world applications, where attack methods evolve rapidly and unpredictably and can be further reinforced through meta-learning and adversarial training strategies aimed at improving generalization to unseen attacks [17]. Nonetheless, self-attention's computational complexity is a major roadblock to deployment on energy-constrained devices. This fact has driven new research into light-weight Transformer alternatives, model compression, sparsity, and efficient attention.
Based on these observations, several specific future research priorities are necessary to address the identified limitations, which are elaborated in the Conclusion and Future Directions section.
Table 3 highlights the strengths, implications, limitations, and gaps identified across recent research:

**Table 3:** Description of Strengths, Implications, Limitations & Gaps Identified

| Aspects | Description | References |
|---|---|---|
| Strength | -Synthesizes recent studies showing clear performance improvements with Transformer and self-supervised learning models.<br>- Highlights end-to-end learning advantages.<br>- Confirms robustness on standard benchmarks. | [Chen et al., 2023], [Wang et al., 2022] |
| Implications | -Encourages adoption of advanced Transformer architectures for spoofing detection.<br>- Promotes standardized metrics (EER, t-DCF) for benchmarking.<br>- Supports scalability in real-world applications. | [Singh et al., 2023], [Liu et al., 2024] |
| Limitations | -No original experimental work; relies on published data.<br>- Lack of latency and resource use analysis for deployment.<br>- Mostly evaluated on curated datasets; lacks noisy, real-world testing.<br>- Explainability and privacy are insufficiently addressed. | [Zhao et al., 2022], [Kumar et al., 2023] |
| Gaps | -Need for robustness testing across languages, accents, and noise.<br>- Research on lightweight, efficient models for edge devices.<br>- Demand for interpretable (explainable) spoofing detection.<br>- Privacy-preserving techniques exploration [18]. | [Chen et al., 2023], [Liu et al., 2024], [Kumar et al., 2023] |

## 5. Conclusion and future scope

This paper focused on evaluating previously developed voice spoofing detection techniques, tracing the evolution from simple machine learning to the more advanced Transformer-based models. While transformer models offer the best detection performance, selecting the best issues persists because of differing dataset compositions, evaluation criteria, and where the system will be deployed—cloud versus edge devices implementations.
For better and stronger detection systems in the future, it is essential to focus on the creation of lightweight Transformer models that can perform well even on low-resource and IoT devices without sacrificing accuracy. Improving the ability to detect novel, out-of-distribution attacks even with limited training data will make systems more flexible and adaptable to novel threats. Standardizing cross-dataset benchmarks will also ensure robust generalization across languages, accents, environmental noise, and recording conditions. Concurrently, safeguarding the privacy of users by ensuring that no sensitive audio data is stored becomes a crucial priority. Systems for spoof detection can be made more robust and stable through the integration of voice data with other indicators, i.e., face expressions or behavioural patterns, to enhance immunity against sophisticated attacks. Meeting these priorities, in addition to clear research questions like "How to optimize Transformers for low-resource IoT devices without compromising on benchmark-level performance?", will drive the next wave of innovation and ensure that voice security technologies are both effective and realistic for deployment in the field.

## References

[1] J. Wu, Z. Evans, T. Kinnunen, and J. Yamagishi, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," *Proc. Interspeech*, pp. 2037–2041, 2015. https://doi.org/10.21437/Interspeech.2015-462.
[2] M. Hasan, T. Tanaka, and K. Takeda, "Voice spoofing detection using MFCC and SVM," *Int. J. Comput. Sci. Netw. Secur.*, vol. 13, no. 6, pp. 45–51, 2013.
[3] T. Kinnunen *et al.*, "Voice spoofing and countermeasures: A survey," *Speech Commun.*, 2017.
[4] H. Tak, K. Lee, and S. Kim, "RawNet2: End-to-end deep neural network for anti-spoofing," *Proc. IEEE ICASSP*, pp. 1271–1275, 2021.
[5] Baevski Alexei, Zhou Henry, Mohamed Abdelrahman, Auli Michael "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, 2020.
[6] Jung Jee-weon, Heo Hee-Soo, Tak Hemlata, Shim Hye-jin, Chung Joon Son, Lee Bong-Jin, Yu Ha-Jin, Evans Nicholas *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2022. https://doi.org/10.1109/ICASSP43922.2022.9747766.
[7] Y. Zhang, J. Wu, and L. He, "Recurrent neural networks for spoofing speech detection," *IEEE Signal Process. Lett.*, vol. 27, no. 1, pp. 88–92, 2020.
[8] X. Li, F. Wu, and W. Huang, "LSTM-based voice spoof detection system," *Proc. IEEE ICASSP*, pp. 1236–1240, 2021.
[9] S. Jung, J. Lee, and K. Park, "Wav2Vec 2.0 with AASIST for robust spoofing detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 789–799, 2022.

[10] S. Chen, J. Liu, and K. Zhang, "Advances in voice spoofing detection using Transformer-based models," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1234–1245, 2023.
[11] Y. Wang, M. Li, and H. Chen, "A comprehensive review of deep learning techniques for spoofing detection," *J. Signal Process. Syst.*, vol. 98, no. 3, pp. 215–230, 2022.
[12] R. Singh, P. Gupta, and S. Kumar, "Transformer architectures for robust voice anti-spoofing," *Proc. IEEE ICASSP*, pp. 456–460, 2023.
[13] F. Liu, X. Zhao, and M. Zhang, "Standardizing evaluation metrics in voice spoofing detection," *Speech Commun.*, vol. 146, pp. 35–42, 2024.
[14] J. Zhao, L. Huang, and T. Wang, "Challenges in deploying voice spoofing detection: Latency and resource constraints," *IEEE Access*, vol. 10, pp. 56789–56799, 2022.
[15] S. Kumar, D. Patel, and A. Sharma, "Explainability and privacy issues in voice spoof detection," *Int. J. Comput. Speech Lang.*, vol. 37, no. 1, pp. 1–15, 2023.
[16] Y. Zhao, X. Li, Z. Wang, and Y. Li, "A Speech Spoofing Detection Method Combining Multi-Scale Features and Cross-Layer Information," *Information*, vol. 16, no. 3, p. 194, Mar. 2025. [Online]. Available: https://doi.org/10.3390/info16030194.
[17] X. Wang and J. H. L. Hansen, "Towards Improving Synthetic Audio Spoofing Detection Robustness via Meta-Learning and Disentangled Training with Adversarial Examples," *IEEE Access*, vol. 12, pp. 65433-65445, 2024, https://doi.org/10.1109/ACCESS.2024.3421281.
[18] L. Tian, A. Kumar Sahu, A. S. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020. https://doi.org/10.1109/MSP.2020.2975749.
[19] H. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, "Audio deepfake detection with self-supervised WavLM and multi-fusion attentive classifier," Dec. 2023. [Online]. Available: https://arxiv.org/abs/2312.08089.