International Journal of Safety and Security Engineering
Vol. 16, No. 7, July, 2026, pp. 1627-1644
Journal homepage: http://iieta.org/journals/ijsse

Audio Deepfake Detection: A Comprehensive Survey of Generation, Datasets, and Detection Strategies

Samar S. Abbas*, Salah Al-Obaidi, Ali Y. Al-Sultan

Department of Computer Science, College of Science for Women, University of Babylon, Hillah 51002, Iraq

Corresponding Author Email: Scw383.samar.sameer@student.uobabylon.edu.iq

https://doi.org/10.18280/ijsse.160718

Received: 15 May 2026
Revised: 6 July 2026
Accepted: 13 July 2026
Available online: 31 July 2026

Keywords:
ASVspoof challenge, deep learning, deepfake, hybrid deep learning model, machine learning, self-supervised learning, speech synthesis, voice conversion

ABSTRACT

Deepfake audio has emerged as a transformative and controversial AI application. This application uses advanced neural networks like transformers and generative adversarial networks (GANs) to produce sounds that resemble human voices. The substantial progress in deep learning (DL), abundant speech datasets, and demand for voice synthesis have all contributed to the development of this technology. Although deepfake audio offers benefits across various sectors, its misuse raises ethical, security, social, economic, and technological challenges. This study introduces a thorough overview of the fundamentals of deepfake audio creation, emphasizing two common spoofing methods: text-to-speech (TTS) and voice conversion (VC). This paper reviews a wide range of deepfake audio detection frameworks, ranging from traditional machine learning (ML) methods and modern DL architectures, including convolutional neural networks (CNNs), long short-term memory (LSTM), transformers, and self-supervised models, to hybrid models. Despite significant advancements, the analysis shows that there are still major issues. These include poor performance in realistic audio environments, a lack of multilingual diversity, and vulnerability to adversarial attacks. In addition, key benchmark datasets are presented as part of the review, and they are important for training and validating detection models. The survey shows that there is a need for robust and reliable detection systems that are transparent and generalizable to ensure digital trust in the era of synthetic media.

1. INTRODUCTION

Recent advances in artificial intelligence, particularly deep learning (DL), have enabled the creation of highly realistic synthetic media, including images, videos, and audio, as well as the manipulation of existing content, leading to the emergence of the term "deepfake" [1-3]. This concept first appeared in 2017, starting with face-swapping in videos and evolving to include lip-syncing and full-face reenactment [4]. Despite the beneficial applications of deepfake technologies in areas such as entertainment, media production, and educational content generation, their misuse poses serious threats to cybersecurity and society. These threats include the spread of misinformation, identity theft, manipulation of public opinion, and interference in electoral processes [5].

In the audio domain, deepfake technology refers to the modification or synthesis of speech. It can alter original statements or create entirely new audio attributed to a given individual. There are several approaches to generating fake audio, including text-to-speech (TTS) synthesis, voice conversion (VC), or voice cloning [6]. Voice cloning is the process of generating synthetic speech that closely resembles the vocal characteristics of a target individual. It is usually done by applying DL models trained on that person's speech data [7]. Specifically, TTS generates synthetic speech directly from text, whereas VC modifies the source speaker's voice to resemble that of a target speaker. These technologies can preserve linguistic content, whereas non-linguistic features such as pitch, timbre, and speech style are altered [8-10].

Synthetic speech generated using deepfake audio technology allows for highly realistic voice reproduction that can closely replicate an individual's voice using limited audio data [11, 12]. Despite their ability to create convincing voice imitations, audio deepfakes present serious security, privacy, and trust risks. Voice-based authentication is increasingly used in security mechanisms across banking, customer service, and smart devices. As a result, synthetic audio can be exploited to deceive these systems, leading to unauthorized access, financial fraud, and identity theft [13, 14]. Moreover, the unauthorized replication of individuals' voices enables impersonation, erodes trust in communication, and poses serious psychological and social risks [15]. The growing realism of audio deepfakes further makes it difficult to trust audio recordings. This issue complicates legal verification and forensic analysis by challenging the authenticity of voice-based evidence used in courts and law enforcement. Therefore, audio deepfake detection is crucial for mitigating these risks [16, 17].

Audio deepfake detection remains a complex and persistent challenge for both humans and automated systems. Empirical studies show that humans struggle to differentiate authentic speech from AI-generated audio. Even with training, accuracy rarely exceeds 73%, especially when the AI-generated speech mimics natural voice [18, 19]. Automated systems also fail to

differentiate between real-world scenarios due to background noise, variable recording quality, and linguistic diversity (e.g., dialects) [20]. This challenge is exacerbated by the rapid advancement of deepfake generation technologies, which often outpace the development of robust detection methods, further diminishing their long-term reliability and effectiveness [21, 22].

This survey provides a broader and more comprehensive review of the field of audio deepfake detection than existing surveys on the topic. In addition to covering conventional machine learning (ML) and DL approaches, this review also analyzes a hybrid detection framework that combines ML, DL, feature fusion, ensemble methods, and transformer-based architectures. Furthermore, this survey expands dataset coverage by examining both established benchmarks and emerging real-world and multilingual resources, including ASVspoof5, Multilanguage Audio Anti-Spoofing Dataset (MLAAD v8), in-the-wild datasets, and the Singing Voice Deepfake Detection challenges. Moreover, it places particular emphasis on recent developments in self-supervised learning (SSL) models such as Wav2Vec2.0 and WavLM, transformer and Conformer architectures, explainable detection systems, adversarial attack resilience, multilingual robustness, and cross-dataset generalization. Reviewing the aspects related to deepfake detection, including deepfake generation techniques, benchmark datasets, ML approaches, DL architectures, hybrid frameworks, and open research challenges within a unified framework, in a single survey article offers a more comprehensive and up-to-date perspective on the development of robust and practical audio deepfake detection systems.

This survey aims to address the aforementioned gaps by offering the following contributions:

- A comprehensive overview covering TTS systems, VC techniques, replay-based attacks, and adversarial strategies for audio deepfake generation methods.
- A systematic examination of audio deepfake detection approaches, including traditional ML, DL, and hybrid models.
- A comparative analysis of widely used benchmark datasets, highlighting their properties, constraints, and suitability for real-world evaluation scenarios.
- Identify major challenges and research gaps, including generalization, robustness, adversarial attack resistance, multilingualism, and model explainability.
- Explore research directions for robust and practical audio deepfake detection systems.

The focus of this study is to examine audio deepfake generation and detection techniques. Furthermore, the associated challenges of identifying synthetic speech in real-world settings will be discussed. Other forms of deepfakes, including visual deepfakes (such as manipulated images and videos) and text-based synthetic content, are outside the scope of this work. Moreover, the primary emphasis of this survey is on technical and methodological aspects of audio deepfake detection, without considering legal, ethical, and social implications except where relevant.

The remainder of this paper is organized as follows. Section 2 presents the main techniques used for audio deepfake generation. Section 3 reviews the commonly used datasets for audio deepfake research, highlighting their characteristics and limitations. Section 4 discusses existing detection methods based on traditional ML. Section 5 explores DL-based detection approaches. Section 6 discusses a hybrid detection framework that integrates multiple architectures. Section 7 identifies the primary challenges and research gaps in the field. Finally, Section 8 concludes the paper and outlines future research directions.

2. AUDIO DEEPFAKE GENERATION

Deepfake audio refers to AI-modified recordings that still retain a natural and convincing sound. Detecting such audio is crucial due to its increasing role in criminal activities recently. Broadly, audio deepfake techniques can be categorized into three main types: speech synthesis, VC, and adversarial attacks [10, 21].

2.1 Text-to-speech

TTS is the technology that converts written text into a speech waveform using linguistic processing, acoustic modeling, and waveform generation. Modern TTS systems use deep neural network (DNN) technologies to produce very realistic, human-sounding audio that significantly improves the overall quality of synthetic voices [23, 24].

A TTS system consists primarily of two parts. These are the front-end and back-end components. TTS front-end components perform the following functions: word segmentation, part-of-speech (POS) tagging, and grapheme-to-phoneme conversion. While back-end components predict prosodic and acoustic representations and convert them into a speech waveform using a vocoder. Developing both parts requires considerable expertise due to numerous complex design decisions [25, 26].

The uses of TTS technology are extensive: some examples include voice assistants [27], navigation devices and systems [28], speech-to-speech translation [29], and educational applications [30]. In addition to its benefits, TTS poses significant risks of misuse. For example, deepfakes or forged recordings can be used to falsely associate audio with a particular person [31]. The synthesis of deepfake audio from TTS involves two key steps: the collection of clean audio or adapting speech data; its associated transcript; and training the TTS system to produce synthetic output with high quality and realism [6].

2.2 Voice conversion

VC is the process of altering a source speaker's voice to resemble that of a target speaker while preserving the original linguistic content [32, 33]. VC employs a variety of speech processing techniques, such as analysis, spectral modification, and speaker characterization, to transform voices. VC techniques have progressed from statistical approaches such as Gaussian mixture models to modern DL architectures, including generative adversarial networks (GANs) and variational autoencoders (VAEs), which leverage large datasets to generate more realistic and flexible synthetic voices. These improvements have made it possible to generate highly natural, human-like voices that closely match the target speaker's identity [34-36].

However, VC poses a significant risk for applications that rely on speaker identity verification, as it can undermine the reliability of authentication systems [37]. These technologies may be increasingly misused to spread disinformation, enable political manipulation, and facilitate economic fraud, making it critical to develop advanced detection methods to identify and mitigate such attacks [38].

2.3 Adversarial attacks

Deepfake detection models and automatic speaker verification (ASV) systems remain highly vulnerable to adversarial attacks. Those attackers use optimization-based techniques to subtly alter audio signals to create perturbations imperceptible to human listeners or might sound like real background noise. These attacks mislead detection and verification models, which means increasing false acceptance and alarm rates, thus compromising system reliability [39-41].

Several specialized adversarial techniques have been developed to exploit the vulnerabilities in speaker verification systems. Malafide is a universal attack that introduces convolutional noise via an optimized linear time-invariant (LTI) filter, reducing the effectiveness of countermeasures while preserving both the quality of speech and the characteristics of the speaker's voice [42]. Malacopula, using a generalized Hammerstein model, manipulates speech signals to exploit ASV vulnerabilities for spoofing [41]. Compared to earlier approaches, this model offers enhanced nonlinear control over amplitude, phase, and frequency components, improving the effectiveness of adversarial perturbations.

The Deep4SNet, which achieves an accuracy of 98.5% in deepfake detection, is highly vulnerable to adversarial attacks. Targeted adversarial perturbations, for example, drastically reduce the detector's performance to nearly 0%, demonstrating the need for more resilient detection methods [43].

3. DATASETS

Datasets are essential for training, validating, and benchmarking detection algorithms, making them fundamental to the progress of audio deepfake detection. Researchers use high-quality datasets to evaluate how well their methods perform across different types of forgery, languages, and recording environments [31]. In the following, we will review and discuss a set of datasets commonly used for audio deepfake detection that were generated using the aforementioned techniques.

(1) ASVspoof2015 [44]

This dataset marks the first major international competition dedicated to exposing vulnerabilities in ASV systems to spoofing attacks and advancing the development of effective countermeasures. Built upon a standardized database, it contains both genuine and spoofed speech samples.

The genuine recordings, collected from 106 human speakers (45 male and 61 female) without any alterations, exhibit minimal channel or background noise interference. In contrast, the spoofed speech was generated by applying various speech synthesis (SS) and VC algorithms to the original recordings. The resulting synthetic copies, generated using multiple forgery algorithms, were systematically partitioned into training, development, and evaluation subsets.

(2) ASVspoof2017 [45]

The primary objective of the ASVspoof2017 competition is to assess how effectively ASV systems can detect spoofing attacks under realistic conditions. By pursuing this goal, the challenge aims to advance the development of a robust framework to combat voice spoofing, with particular emphasis on detecting replay attacks.

The ASVspoof2017 database contains 3,565 genuine utterances from 42 speakers, collected through the Red Dots program, and 14,465 spoofed audio samples from 177 speakers, created by replaying the original recordings across a wide range of devices and acoustic environments to simulate realistic replay scenarios.

(3) ASVspoof2019 [46]

The edition of ASVspoof2019 was the first major evaluation campaign to deal with all three types of spoofing attacks: TTS, VC, and replay attacks—within a unified framework. It's the third expansion version of the ASVspoof challenge series compared with ASVspoof2015 and ASVspoof2017. This dataset is separated into two scenarios: logical access (LA) and physical access (PA). The LA scenario emphasizes the use of advanced synthesis and VC technologies such as neural acoustic and waveform models, whereas the PA scenario addresses replay spoofing attacks in simulated physical environments. The 2019 edition also introduces a new scale for reliability, incorporating the impact of spoofing and countermeasures. Original recordings were obtained from 107 speakers, 46 male and 61 female under conditions with minimal channel distortion and background noise. Systematically, spoofed speech samples were then generated from this source material using a variety of spoofing algorithms to achieve wide coverage of potential attack strategies.

(4) ASVspoof2021 [47]

ASVspoof2021 is the fourth edition of the ASVspoof challenge series. It was designed to evaluate the detection of synthetic and manipulated speech across three evaluation partitions (LA, PA, and DF). In this edition, no new training or development datasets were provided; participants relied on the ASVspoof datasets previously released to develop their systems.

The only difference in this evaluation set compared to past evaluations is the added emphasis on differences across channels that produce varying degrees of coding and transmission artifacts (i.e., telephony networks and IP-based communications).

The evaluation dataset for both the LA and PA tasks uses recordings taken from the same 48 speakers that were used for evaluation in the ASVspoof2019 challenge. In addition, the Deepfake task dataset was collected from VCTK, along with additional unpublished datasets, to provide a wider variety and greater complexity in fake speech scenarios while also yielding a more realistic overall evaluation. The evaluation dataset presents completely new and highly variable challenge conditions against which to evaluate current state-of-the-art spoofing detection systems.

(5) ASVspoof5 [48]

ASVspoof5 is the fifth edition of ASVspoof, which is built to assess the effectiveness of systems in detecting spoofing and deepfake audio. This dataset represents a major shift toward "in-the-wild," which means that the recording is done in a crowdsourced environment instead of studio-quality recordings. It is built upon the Multilingual Libri Speech (MLS) English partition, offering diverse sources and a broad range of speaker voices using diverse consumer devices in uncontrolled acoustic environments.

ASVspoof5 expands the challenge by incorporating a suite of 32 different attack algorithms that include legacy and advanced TTS, VC, and adversarial attacks for the first time. These additions aim to generate synthetic samples that are more sophisticated and difficult to detect. The dataset includes seven speaker-disjoint subsets to support tasks ranging from standalone detection to spoofing-robust ASV. Furthermore, it

provides around 30000 additional speakers for training robust encoders.

(6) Fake or Real [49]

This open-source dataset supports synthetic speech detection research. It contains 198,000 English-language audio samples designed to train DL models without overfitting. It includes 87,000 synthetic utterance audio recordings generated by Deep Voice 3, Amazon AWS Polly, and Google Cloud TTS, and 111,000 real utterance audio recordings created from the Arctic Dataset, LibriSpeech, VOXForge, and YouTube educational videos featuring diverse speakers. The dataset is available in four versions (FoR-original, FoR-Norm, FoR-2sec, and FoR-rerec).

(7) Wavefake [50]

This dataset comprises 117,985 generated audio files (16-bit PCM WAV) totaling approximately 196 hours. Primarily based on the LJSpeech (English) and JSUT (Japanese) language sets, the samples were generated using six advanced architectures (MelGAN, MelGAN Large, FB-MelGAN, MB-MelGAN, HiFi-GAN, Parallel WaveGAN, and WaveGlow) and cover ten subsets, including two languages and a complete TTS script with new phrases.

(8) ADD2022 [51]

This audio deepfake detection challenge addresses gaps left by previous challenges, such as ASVspoof2021, by focusing on realistic and complex scenarios. It covers audio generation and detection through three main tracks: Low-Quality (LF) fake audio detection, Partial Fake (PF) detection, and Fake Game (FG). Using diverse datasets such as AISHELL-3, the challenge employed well-defined protocols and evaluation metrics. The results highlight significant difficulties in model generalization and the need for improved evaluation metrics.

(9) In-the-Wild [52]

The in-the-wild dataset is a widely used benchmark specifically generated to address the generalization gap in detecting speech deepfakes. Instead of using controlled laboratory settings to generate datasets, this dataset consists of real-world audio collected from publicly available sources, such as social media and video streaming platforms. It includes both genuine and synthetic audio recordings, totaling approximately 38 hours, with 20.8 hours of authentic speech and 17.2 hours of deepfake audio, produced by 54 English speakers. Due to generating the samples from various unknown generating algorithms and recording them under diverse acoustic conditions with varying compression levels, it serves as a rigorous test for how well a detection model can perform against realistic, unseen threats.

(10) Multilanguage Audio Anti-Spoofing Dataset [53]

MLAAD v8 addresses linguistic bias in existing deepfake detection datasets by providing a multilingual corpus. It is built as an extension of the M-AILABS Speech Dataset, whose original bona fide recordings include eight languages (English, French, German, Italian, Polish, Russian, Spanish, and Ukrainian). These original recordings consist of authentic human speech collected from audiobooks, public speeches, and interviews. By using the original or translated transcripts as input to TTS systems, MLAAD v8 extends these source recordings by generating 570.3 hours of synthetic audio across 40 different languages, using 119 advanced TTS models representing 58 distinct architectures. MLAAD v8 offers a valuable resource for evaluating cross-lingual generalization and real-world performance of audio anti-spoofing systems by incorporating an extensive collection of synthetic voices and languages.

(11) The Singing Voice Deepfake Detection challenge [54]

The singing voice deepfake detection **(**SVDD) is the first research challenge to develop systems that can identify the authenticity of human-sung audio recordings versus AI-made recordings of the same singer. A recent increase in advancements with generative models using AI technology to replicate the human voice for musical purposes has made this research topic significantly more relevant. The SVDD project hopes to catalyze additional research into how we can prevent the infringement of an artist's right to their own work and create systems to automate this process. This cannot be accomplished without sophisticated forms of DL, which can extract information from both natural and artificial sounds that an investigator can use when determining the authenticity of the audio source.

To aid in research in this area, the SVDD2024 challenges created two benchmark datasets: CtrSVDD and WildSVDD. CtrSVDD consists of pure singing recordings without any accompanying music, as well as some synthetic samples produced by using a variety of advanced audio transformation and synthesis techniques. On the other hand, WildSVDD attempts to simulate real-world conditions as closely as possible by including samples taken from social media and including both types of recordings (real and fake), each with its own varying degrees of background music and different recording environments (e.g., sound quality).

Table 1. Comparative summary of prominent datasets for synthetic speech, voice conversion (VC), and replay attack detection sorted by year of creation

Dataset	Spoofed Type	Format	Sample Rate	Number of Bona Fide Samples	Number of Spoofed Samples
ASVspoof2015	TTS, VC	Flac	16 KHz	16651	246500
ASvspoof2017	Replay	Wav	16 KHz	3565	14465
ASVspoof2019	TTS, VC	Flac	16 KHz	12483	108978
ASVspoof2021	TTS, VC	Flac	multiple	18452	163114
ASVspoof5	TTS, VC, adversarial attack	Flac	16 KHz	188819	815262
FoR 2019	TTS	Wav	multiple	111000	87000
ADD2022	TTS, VC	Wav	16 KHz	3012	24072
in-the-wild2022	Not provided	Wav	16 KHz	19963	11816
MLAAD2024	TTS	Wav	22 KHz	0	76000

Note: TTS = text-to-speech, VC = voice conversion.

In Table 1, we summarize the core characteristics of each dataset, including the static information, and report the accuracy results for each dataset based on the results reported in the studies that make up the basis of this review. This table also lists the various types of attacks represented within each dataset, e.g., TTS attacks and VC attacks. In addition, in this table, we found information about more recent adversarial attack types and some technical characteristics. Furthermore, we demonstrate an increase in the volume of data available to researchers by showing an increase in the number of spoofed samples within each dataset. At the same time, we show that there has been a significant increase in both the complexity and resource requirements of the modern-day detection models across datasets over time.

4. MACHINE LEARNING-BASED AUDIO DEEPFAKE DETECTION

ML was among the first systematic methods for detecting audio deepfakes. Early ML-based audio deepfake detection methods used handcrafted acoustic feature engineering combined with traditional classifiers. These methods aimed to exploit measurable statistical discrepancies between real and synthetic speech, particularly within benchmark datasets such as ASVspoof. Although these models are often outperformed by DL, traditional ML frameworks provide valuable baselines and interpretable insights into acoustic cues.

4.1 Feature engineering for spoof detection

The manually designed acoustic features are the building blocks of the traditional audio deepfake recognition system. These features try to detect spectral, temporal, phase, and statistical discrepancies between genuine and manipulated speech. These include Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), spectral roll-off, spectral centroid, spectral contrast, spectral bandwidth, chroma features, zero-crossing rate (ZCR), and phase-based features.

Early studies have proven that the phase can offer discriminative characteristics that cannot be captured by magnitude-based features. Using the cos-phase and modified group delay function in Gaussian Mixture Models resulted in better equal error rates (EER) compared to the methods employing MFCC features, implying the presence of phase inconsistency for detecting artifacts during voice conversion [55]. The success of such feature representation was mainly tested on voice-conversion scenarios, which means generalizing the results to modern audio deepfake is challenging.

Benchmark tests following that have proven that the performance of the classifiers is significantly dependent upon the relationship between the acoustic representation, the applied normalization method, and the kind of spoofing attack used. Comparison of the GLDS-SVM, GMM-SVM, GMM-ML, GMM-UBM, and *i*-vector systems based on ASVspoof2015 proved that some classifiers were better at dealing with development data, while generative GMM classifiers gave better results with the evaluation data, including unknown attacks [56]. The poor performance of *i*-vectors was notably improved after applying the within-class covariance normalization, thus proving that preprocessing and normalization could be highly influential for the robustness of handcrafted systems. Moreover, it could be concluded that synthetic speech was easier to detect than voice-converted speech since the conversion technique could retain some properties of the speaker's voice.

Another advantage of the multi-feature anti-spoofing systems was related to the use of both phase and magnitude components of the speech signal. The anti-spoofing systems that included descriptors such as noise-unaware local binary pattern, modified group delay, and cosine normalized phase had lower error rates than traditional GMM likelihood-based approaches [57].

All these works serve to demonstrate the progression of spoof detection technology, from being based on isolated phase-aware features to incorporating features from magnitude-phase representation [55-57]. Yet, the effectiveness of all of them depends on manual feature selection, classifier design, score-level fusion and normalization, and the similarity of the test attack to the training one.

4.2 Classical machine learning classifiers

In terms of discrimination of artificial speech from authentic audio data, classical ML classifiers created one of the earliest attempts to develop structural solutions for differentiation between real voice and imitating or synthetically generated audio. The performance provided by the authors illustrates the potential of relatively simple ML algorithms when used in conjunction with highly discriminative statistical descriptors.

For instance, acoustic features computed based on entropy in combination with a logistic regression model yielded 98% accuracy of distinction between authentic and artificial speech [58]. This example proves that even relatively simple models can work efficiently as long as statistical descriptors provide sufficient differences between real and artificial data.

Support vector machines have demonstrated remarkable performance in audio spoofing detection using binary classification as well. A quadratic SVM provided a highly accurate performance while classifying real audio from artificially synthesized audio, surpassing other linear and distance-based classifiers [59]. Such results prove the usefulness of non-linear boundaries in separating handcrafted representations. Yet, the use of an internal dataset as a benchmark limits the generalization of the findings. Moreover, the performance of SVM highly depends on the choice of feature space and acoustic conditions represented in the dataset used for training.

Another strand of research concerns the question of how effective the combination of short- and long-term features could be. Prediction-based features based on linear auto-regression modeling of the temporal structure of audio were proposed to separate genuine from computer-generated speech [60]. Testing on ASVspoof2019 proved better performance than several baselines, including voice synthesis using DL techniques. Thus, the proposed method proved to detect artifacts that are not perceptible to humans in handcrafted prediction features. Yet, the dependence on the predefined linear prediction features could limit the flexibility of this technique for other synthetic speech generation methods, codecs, languages, and recording conditions.

Other comparisons between different classical classifiers have also demonstrated that not only the choice of the algorithm matters but also feature dimensionality and redundancy. The framework that uses MFCC, spectral roll-off,

spectral centroid, spectral contrast, spectral bandwidth, and ZCR to evaluate the performances of SVM, multilayer perceptron, decision tree, logistic regression, Naive Bayes, and XGBoost classifiers on TTS-generated audio [61] serves as an example. Due to the extensive nature of the generated feature space, PCA was needed to keep only the most informative descriptors and eliminate redundancy. SVM produced the best results in the comparison and demonstrated accuracies of 97.5% and 98.83% on the FOR2sec and FoR-rerec datasets, respectively [61]. While these results demonstrate the ability of classical classifiers to produce high accuracy when combined with well-tuned features, the need for dimensionality reduction demonstrates that it is inefficient to create a big set of handcrafted features in the absence of automatic feature representation learning.

Therefore, it can be stated that classical ML approaches are computationally efficient and rather interpretable solutions for the problem of audio deepfake detection. However, their efficiency is often linked to specific data and controlled conditions and to handcrafted feature optimization.

4.3 Ensemble and explainable approaches

Conventional ML research has progressively moved away from individual binary classifiers to approaches aimed at enhancing generalization capabilities, integration of complementary decision-making schemes, and the interpretation of model decisions.

One-class classification is among the first approaches proposed to deal with the issue of unexpected spoofing attacks. The spectral-temporal textures using local binary patterns have been integrated with the one-class approach, where training has been done solely on speech data [62]. Unlike traditional binary classification models trained on real and spoofed samples, this approach focuses on capturing the distribution of real speech and the deviation from the distribution. The obtained results showcased the effectiveness of one-class learning for detection of various types of spoofing without having examples of each type of attack. On the other hand, the natural speech could vary greatly across speakers, languages, communication channels, and even recording devices. In this case, a one-class classifier might either mistakenly classify variations as spoofing or simply fail if the generated speech resembles the distribution.

Explainability has also been introduced in a handcrafted ML approach to identify the acoustic features responsible for the classification process. A multi-dataset framework for extraction of MFCC, spectral, RMS energy, ZCR, and chroma-based features and evaluation of XGBoost, random forest, k-nearest neighbors, and multilayer perceptron classifiers has been used. Accuracies reached 89%, 94.5%, and 94.67% on ASVspoof2019, FakeAVCelebV2, and in-the-wild audio deepfake datasets, respectively. The next step was the identification of the most significant features by using SHapley Additive explanations [63]. It also allows more transparency than strictly performance-driven classifiers and facilitates forensic analysis. Still, the explanations would be limited to the predefined set of features fed into the model. Hence, SHAP would be able to explain how the classifier uses the selected set of descriptors but would not be able to account for the acoustic information missing from the handcrafted feature set.

Tree-based ensembles have also been utilized in order to achieve better nonlinear classification with relatively modest computational expenses. The evaluation of mel-spectrograms with random forest, gradient boosting, and XGBoost classifiers revealed that the XGBoost classifier reached the highest performance level with 99.32% accuracy on the DEEP-VOICE dataset with five-fold cross-validation [64]. Thus, it is evident that the boosted trees were capable of learning non-linear decision boundaries from time-frequency representations. However, the lack of diversity in the dataset and controlled environment limits the generalizability of the results for real-world audio, unseen speakers, compression, channel distortions, and generation models.

In sum, one-class, explainable, and ensemble ML methods tackle various problems associated with traditional classification methods. One-class ML algorithms aim at better detection of novel attacks, explainable approaches provide more clarity regarding the role of each acoustic feature, and ensemble algorithms enhance classification due to the use of many decision trees [62-64]. However, these solutions cannot be generalized across datasets and generation approaches. Moreover, explainability is limited by the level of the handcrafted representation of data, and the accuracy of ensemble algorithms can degrade significantly in real-world scenarios.

4.4 Strengths and limitations of machine learning approaches

Conventional deepfake detection systems based on ML algorithms are effective in terms of computation but not in terms of interpretability or applicability in forensic/low resource settings. Since these systems can use engineered features, they provide useful information about the acoustic cues that matter most, allowing them to be effectively utilized for explainable AI systems.

Despite the advantages of these models, there are significant drawbacks. The extracted features are limited in terms of their relevance and how well they can represent the actual characteristics of an object or scene. These limitations may be especially pronounced for features that were generated from more complex objects and scenes, such as modern generative models, as the inability to accurately represent complex features may limit the model's ability to accurately predict or detect objects. Additionally, while there may be high performance on certain datasets, this may not translate to other datasets or languages. The features created by hand will also be overly sensitive to channel conditions, noise, and variations in the recording environment, leading to less accuracy when used in real-world applications. Lastly, traditional feature extraction methods will struggle to keep up with the ever-evolving subtlety of acoustic differences created by newer generative models, especially those based on neural codecs or large audio-language models.

To sum up, ML-based methods provide the basis for audio deepfake detection and remain valuable in specific situations. However, their scalability and adaptability are limited to modern DL and SSL models. Table 2 provides a summary of various research studies focused on the detection of spoofed and AI-generated speech using conventional ML and discriminative-based methods.

Table 2. Comparative analysis of machine learning (ML) approaches for synthetic speech and spoofing detection

References	Method	Dataset	Results	Advantage	Limitation
Wu et al. [55]	GMM	NIST 2006	EER = 2.35%	Superiority of phase-based features. Better spoof discrimination.	Limited to VC attacks. Noise-sensitive and GMM-dependent.
Hanilçi et al. [56]	GLDS/GMM-SVM	ASVspoof2015	EER = 3.01%	Better discrimination than GMM baselines. Strong benchmark baseline.	Weak against unseen attacks. Depends on handcrafted features.
Liu et al. [57]	ASSV, SVM	ASVspoof2015	EER = 3.086%	Magnitude and phase outperform single features. Improved robustness via score fusion.	Limited generalization. Manual feature engineering.
Rodríguez-Ortega et al. [58]	LR	H-Voice	Accuracy = 98%	Lightweight and interpretable. Competitive with simple features.	Manual feature extraction. Limited scalability.
Singh et al. [59]	Q-SVM	Internal	Accuracy = 97.56%	Simple and computationally efficient. Suitable for forensic use.	Non-benchmark dataset. Less adaptable than DL.
Borrelli et al. [60]	SVM, RF	ASVspoof2019	Accuracy = 93%	Highly interpretable. Works across multiple detection scenarios.	Performance drops in noisy environments. Limited automatic feature learning.
Iqbal et al. [61]	SVM, XGB, MLP	FoR	Accuracy = 98.83%	High accuracy with low computational cost. Efficient feature selection.	Single dataset evaluation. Feature dependent.
Alegre et al. [62]	One-Class SVM, *i*-vector	NIST SRE	EER = 7.6%	Better detection of unseen attacks. Requires only genuine speech.	Older datasets. Less effective against modern deepfakes.
Bisogni et al. [63]	RF, XGB, MLP	in-the-wild	Accuracy = 94.67%	Explainable and computationally efficient. Suitable for practical deployment.	Sensitive to feature quality. Lower robustness than SSL models.
Bohara and Bairwa [64]	XGB, RF, GB	Deep voice	Accuracy = 99.32%	High accuracy with low computation. Easy to interpret.	Small dataset. Limited real-world validation.

Note: The performance metrics reported in these tables are taken from the original studies and should not be interpreted as directly comparable rankings. This is because the reviewed studies differ in terms of datasets, class distributions, evaluation protocols, and reported metrics. Therefore, these values are presented only as descriptive summaries of methodological approaches and reported evaluation outcomes.
EER = equal error rate, SVM = support vector machine, LR = Logistic Regression.

5. DEEP LEARNING-BASED AUDIO DEEPFAKE DETECTION

DL-based audio deepfake detection leverages learning representations from speech automatically using advanced neural networks. The most popular approaches used in this field are convolutional neural networks (CNNs), long short-term memory (LSTM), and transformers. A CNN may capture both frequency patterns (spectral) as well as time-frequency patterns, while LSTM captures long-term temporal dependencies present in audio sequences [65, 66]. In recent years, transformer and self-supervised models have shown improved robustness, scalability, and generalizability against many new deepfake audio generation methods.

5.1 Convolutional neural network-based spectral and time-frequency modeling

Initial DL methods for audio spoofing detection mainly relied on CNNs integrated with handcrafted or time-frequency acoustic features. Such as MFCC, CQCC, and spectrograms. The early frameworks used standard CNNs, residual CNNs, light convolutional neural networks (LCNNs), and transfer-learning-based ResNets to capture discriminative artifacts associated with spoofed speech. Residual CNNs showed the possibility of leveraging multiple representations of the acoustic signal for spoofing detection. The proposed ResNet framework utilizing MFCC, CQCC, and spectrogram as inputs yielded a low EER on the logical-access subset of the ASVspoof2019 dataset, indicating the effectiveness of deep residual learning in discriminating between real and synthetic speech [67]. The downside of using multiple representations is a higher computational burden and no performance guarantees when switching to cross-dataset evaluation. Subsequent research investigated lightweight architectures to reduce the computational complexity associated with standard residual networks. Combining an LCNN with a genuinization Transformer resulted in an EER of 4.070% on ASVspoof2019 [68]. This result suggests that computational efficiency can be improved while preserving the ability to learn discriminative spoofing patterns. However, the evaluation was limited to logical-access attacks within a single benchmark; therefore, robustness to physical-access attacks, noisy audio, and real-world conditions remains unknown.

Transfer learning has been applied for improving representation learning while mitigating optimization challenges for deep architectures. ResNet-34 models have been able to benefit from residual connections that alleviated vanishing gradients and the capability to learn multi-scale acoustic features [69]. However, despite the abovementioned benefits, such models are computationally expensive and require diverse training data to support reliable generalization.

A more localized direction employs frequency-domain CNNs to analyze spectral abnormalities without modeling the

complete temporal sequence. This approach achieved 85.99% accuracy on ASVspoof2019, indicating that frequency-domain features can be used to distinguish between genuine and spoofed speech [70]. However, its lower performance relative to more advanced models shows that frequency-domain approaches are incapable of modeling long-term temporal and context-related inconsistencies.

In summary, it can be said that CNN, LCNN, and ResNet approaches have demonstrated that deep feature learning is effective for detecting local spectral and time-frequency feature artifacts [67-70]. The major drawback of these methods is that their performance is highly dependent on the selected acoustic front-end and similarity of training and evaluation environments.

5.2 Temporal, contextual, and sequence-modeling architectures

Although convolutional layers have been successful in local pattern detection, they may be limited in capturing long-range temporal relationships and contextual dependencies. This limitation has encouraged the development of transformer-based, Conformer-based, and state-space models that are able to capture interactions within long audio sequences. Hierarchical Conformer models integrate convolutional modeling of local speech patterns with the use of transformers for global sequence modeling. The use of hierarchical pooling and multi-class tokens helps in the integration of the spoofing information on multiple scales [71]. This design produces more sophisticated sequence representations than the traditional CNN models; although its EER suggests that architectural sophistication alone is insufficient to guarantee strong detection performance. The complexity of the model and the training and dataset characteristics remain important factors.

The integration of Conformer back-ends and Wav2Vec 2.0 representations resulted in significantly lower error rates for speech samples of varying lengths. Such models were able to achieve an EER of less than 1% in logical-access experiments by combining contextual SSL and sequence modeling for accommodating the variability in the audio duration [72]. These results highlight the advantages of using pretrained representations together with architectures that can model simultaneously local and global features.

State-space models provide an alternative to standard transformer architecture due to their ability to work with long sequences without some of the computational limitations associated with attention mechanisms. Using a bidirectional Mamba-based architecture for extracting low-to-high-level speech features via parameterized sinc functions and convolutional blocks, followed by bidirectional sequence modeling, state-space processing is able to capture long dependencies while maintaining competitive detection performance [73]. However, such architectures are more complex to design, and they still need large and diverse training datasets. In general, Conformer and Mamba models provide more advanced temporal modeling than a convolution-only approach [71-73]. Nevertheless, these benefits are associated with higher complexity, higher computational cost, and continued sensitivity to input duration and model configuration.

5.3 Self-supervised speech representation learning

One of the most promising approaches to audio deepfake detection is self-supervised learning (SSL). Unlike other methods that rely solely on hand-crafted front-ends, SSL methods employ pre-trained speech models such as Wav2Vec 2.0, XLS-R, HuBERT, and WavLM for extracting high-level representations from raw audio samples.

Pre-training on authentic speech was found to be beneficial to spoofing detection even when no artificial examples were needed in the initial learning phase. It was also demonstrated that combining Wav2Vec 2.0 representations with an AASIST back-end resulted in lower error rates compared to sinc-based front-ends, which shows that the knowledge learned from authentic speech can support the detection of different spoofing attacks [74]. This means that SSL representations encode transferable speech characteristics rather than only attack-specific acoustic cues.

Numerous evaluations of SSL front-ends and neural back-end classifiers resulted in near-zero EERs in several ASVspoof challenges [75]. These findings indicate that selecting an appropriate representation may be as important as selecting the classifier. But these systems require careful fine-tuning, and their strong performance on benchmark datasets does not necessarily transfer to audio that differs in language, codec, recording device, or generation method.

Architecture search techniques have also been employed together with SSL to reduce dependence on handcrafted detection architectures. Using Wav2Vec 2.0 representations along with Differentiable Architecture Search (DARTS) produced an EER of 1.08% in ASVspoof2019 LA [76]. In this framework, the pre-trained front-end extracts speech embeddings, and the architecture search process identifies a suitable classification architecture with reduced manual design. Though this alleviates the burden of designing the architecture, it can be very computationally intensive, and performance remains dependent on the quality of the pretrained representation.

Late fusion of ensembles of fine-tuned WavLM models shows the increased robustness resulting from combining model predictions [77]. Their usage in the ASVspoof5 challenge demonstrates the effectiveness of SSL ensembles in dealing with diverse attack conditions. But ensembling of models is an additional requirement, which increases memory consumption and latency.

Another technique is provided by Siamese learning to boost discrimination without increasing the size of the model too much. The combination of Wav2Vec features and a Siamese SENet34 model achieved an EER of 1.21%, which was higher than an FF-LCNN baseline that obtained an EER of 2.06% [78]. This result suggests that learning similarities can improve the separation of real and fake representations. However, Siamese systems should be carefully designed and optimized since their performance depends on the selected margins and other training hyperparameters.

Overall, the techniques based on SSL exhibit higher robustness compared to handcrafted features, especially in cases of unknown attacks, compression, codec variation, and cross-dataset differences [74-78]. The main problems of these techniques are expensive pretraining, high resource needs, sensitivity to fine-tuning, and the presence of biases from original pretraining data in the models.

5.4 Attention, multi-branch, and feature-fusion architectures

Another research direction has explored attention mechanisms, multi-branch processing, and feature fusion for

complementary acoustic representations. These models try to boost discrimination ability by utilizing multiple feature spaces, including spectral, temporal, semantic, emotional, and pretrained representations rather than just one.

The combination of three-dimensional CNNs, bidirectional LSTMs, attention mechanisms, and classic classification modules enables the modeling of both local time-frequency structure and high-level emotional or semantic features [79]. The reported AUC of 0.98 shows that complementary emotional signals may help to distinguish synthetic speech. However, the usage of various DL components increases computational costs and may restrict practical scalability.

Multi-feature attention architectures were also trained on the MFCC, LFCC, and Chroma-STFT representations before applying dense classification modules [80]. As a result of the utilization of multiple complementary acoustic views, such neural network architectures show better performance compared to simple CNN and rule-based models. The reported results on the in-the-wild and FoR datasets confirm the effectiveness of feature fusion for various recording conditions. However, the success of feature fusion heavily depends on the choice of features, since the latter must be complementary.

Attention blocks have been added to existing CNN networks to highlight the most informative spectral and temporal regions. VGGish network enhanced by a Convolutional Block Attention Module reached extremely low EER in the ASVspoof2019 contest, which suggests that channel and spatial attention might improve the process of extracting spoofing features [81]. On the other hand, preprocessing intensity and noise sensitivity may significantly decrease the reliability of such systems in practice.

Efficiency-oriented attention systems have tried to combine high feature specificity with relatively low resource usage. A combination of the MobileNetV2 network and multi-frequency attention achieved good accuracy in the FoR-norm dataset by highlighting informative frequency features [82]. However, even though this architecture is less resource-intensive than many transformer-based networks, its evaluation is still limited, and its multilayered structure increases implementation complexity.

Feature-specific ensemble methods proved that the choice of acoustic representation also plays a significant role in the success of the system. Evaluating MFCC, CQCC, Mel-spectrogram, and spectral-centroid representations using pretrained CIFAR10 and ResNet50 architectures revealed that MFCC input data performed best among other variants of the FoR dataset, reaching 99.12% accuracy, 97.54% precision, 98.44% recall, 98.12% F1-score, and 0.88% EER [19]. These findings show the utility of hybridizing pretrained CNNs with discriminative acoustic front-ends. Nevertheless, such an overreliance on one specific dataset poses problems concerning overfitting and cross-domain validity.

Attention-based and fusion models can thus help to enhance discrimination through the combination of various acoustic views as well as focusing on salient parts [19, 79-82]. Limitations of their application are greater complexity, the need for proper feature combinations, costly preprocessing, and a lack of confirmation of reliable behavior in unseen acoustic conditions and under unknown data generation.

5.5 Cross-dataset generalization and robustness to emerging attacks

Despite considerable advances in terms of benchmark performance, generalization to new attacks, new languages, new datasets, codecs, and recording settings is the critical issue in deepfake detection. High performance in a well-known benchmark does not always imply the learning of generic properties of the spoofing process by the model.

It has been demonstrated that direct comparison of handcrafted and learned representations leads to the conclusion that manually designed features can perform very poorly under out-of-domain circumstances. Training a ResNet-18 classifier on ASVspoof2019 and evaluating its performance on ASVspoof2021 DF and in-the-wild subsets yielded EER greater than 50% for handcrafted representations [83]. In contrast, SSL representations, such as those obtained from XLS-R, HuBERT, and WavLM, significantly decreased EER, especially when several learned representations were combined. This comparison proves that learned speech representations are much more transferable than handcrafted descriptors. Nevertheless, the remaining EER clearly shows that SSL does not fully solve the generalization problem.

Knowledge distillation has been studied to reduce the computational complexity of models. Self-distillation approaches leverage fine details provided by the deeper teacher neural network and propagate them to the shallow student model, making the inference faster without sacrificing significant performance gains [84]. Testing on the logical- and physical-access subsets of ASVspoof2019 revealed some gains compared to baseline systems. On the other hand, the findings have been confirmed only on one dataset family, and the performance of the distilled models under practical conditions of mismatch has yet to be explored.

Codec-based audio-language models that emerged rapidly have presented a different threat due to dissimilar artifacts compared to the standard vocoder models and voice converters. To train detection models against codec-based attacks, special codec-aware data collections and training procedures were developed [85]. The codec-aware model reported the EER of 0.616%, outperforming the baseline models and showing that attack-specific training is beneficial for next-generation synthetic speech attacks. At the same time, the speech-oriented architecture and lack of noisy-uncontrolled testing limit the conclusions. Gated convolutional architectures based on XLS-R representations have also been considered for cross-domain and multilingual robustness. Centered kernel alignment-based diversity regularization promotes complementary learning of different branches and avoids redundancy in feature learning [37]. Low EER values on ASVspoof2019 and ASVspoof2021 indicate high multilingual and cross-domain capability. However, the model needs diverse training data and is computationally heavy.

The above results demonstrate that the robust audio deepfake detector requires more than just a large-capacity classifier. Generalization is highly dependent on training diversity, transferable representations, adaptation to new synthesis techniques, multilingual testing, and strategies to prevent domain-specific learning [37, 83-85].

Overall, DL-based audio deepfake detection has evolved from CNN-based spectral modeling toward transformer-style sequence modeling, SSL-based representation learning, attention-driven feature fusion, and generalization-focused training. CNN and ResNet models are effective for detecting local spectral artifacts, while Conformer and Mamba-based architectures improve long-range temporal modeling. SSL representations offer stronger robustness than handcrafted features, and fusion or attention mechanisms can enhance

discriminative learning by combining complementary cues. Despite these advances, the shared limitations remain clear: many systems are still evaluated mainly on benchmark datasets, performance may drop under cross-dataset or real-world conditions, and model robustness depends on training diversity, feature representation, architectural complexity, and unseen attack types. Therefore, the main research gap is the need for audio deepfake detection systems that combine high accuracy, interpretability, computational efficiency, and reliable generalization across unseen datasets, languages, codecs, and synthesis methods.

Table 3. Comparative analysis of recent trends in deep learning (DL)-based audio deepfake detection

References	Method	Dataset	Results	Advantage	Limitation
Alzantot et al. [67]	ResNet	ASVspoof2019	EER = 2.78%	Higher accuracy than ML methods. Eliminates manual feature engineering.	High computational cost. Limited cross-dataset validation.
Wu et al. [68]	FG-LCNN	ASVspoof2019	EER = 4.070%	Lightweight and efficient. Better unseen attack detection.	Only LA evaluated. Single dataset study.
Rahul et al. [69]	ResNet-34	ASVspoof2019	EER = 5.32%	Multi-scale feature learning. Better sequence modeling than CNNs.	Computationally expensive. Limited dataset diversity.
Bartusiak and Delp [70]	CNN	ASVspoof2019	Accuracy = 85.99%	Effective frequency-domain analysis. Simpler than transformer models.	Lower robustness to new attacks. Limited generalization.
Shin et al. [71]	HM-Conformer	ASVspoof2021	EER = 15.71%	Multi-scale feature learning. Better sequence modeling than CNN.	High model complexity. Requires large datasets.
Rosello et al. [72]	Conformer	ASVspoof2021	EER = 0.97%	SSL improves robustness. Handles variable-length audio well.	Computationally intensive. Limited benchmark evaluation.
Chen et al. [73]	RawBMamba	ASVspoof2019/ 2021	EER = 1.19%	Strong cross-dataset generalization. Captures long-range dependencies.	Complex architecture. Requires large training data.
Tak et al. [74]	Wav2vec2 AASIST	ASVspoof2021	EER = 0.82%	Better robustness than handcrafted methods. Learns generalized speech representations.	Expensive pretraining, High resource requirements.
Wang and Yamagishi [75]	BiLSTM/LCNN	ASVspoof2015/ 2019/2021	EER = 0.17%	Strong performance across datasets. Effective SSL feature extraction.	Resource-intensive training. Requires careful fine-tuning.
Wang et al. [76]	Wav2Vec2.0 light DARTs	ASVspoof2019/ 2021	EER = 1.08%	Automated architecture optimization. Better than manually designed models.	High search cost. Sensitive to pretrained features.
Combei et al. [77]	WavLM fusion	ASVspoof5	EER = 17.08%	Robust SSL representation. Ensemble improves stability.	High inference cost. Depends on augmented data.
Xie et al. [78]	Siamese MLP	ASVspoof2019	EER = 1.21%	Better unseen attack detection. Parameter-efficient.	Complex optimization. Sensitive hyperparameters.
Conti et al. [79]	3D-CRNN, BiLSTM, Attention RF	ASVspoof2019	AUC = 0.98	Uses complementary emotional cues. Good generalization.	High complexity. Limited attack coverage.
Krishnan et al. [80]	MFAAN	in-the-wild FoR	Accuracy = 98.93%	Multi-feature fusion improves accuracy. Strong real-world performance.	Risk of overfitting. Requires large datasets.
Kanwal et al. [81]	VGGish-CBAM	ASVspoof2019	EER = 0.07%	Attention improves feature learning. Excellent benchmark accuracy.	Expensive preprocessing. Noise-sensitive.
Feng [82]	MFCMNet	FoR-norm	Accuracy = 88.2%	Fine-grained feature extraction. Efficient attention mechanism.	Limited benchmark validation. Complex architecture.
Kaur et al. [19]	ResNet50	FoR	Accuracy = 99.12%	Very high detection accuracy. Ensemble improves robustness.	Dataset dependent. High computational cost.
Yang et al. [83]	ResNet18	ASVspoof2019/ 2021 in-the-wild	EER = 24.27%	SSL outperforms handcrafted features. Feature fusion improves robustness.	EER remains high. High computational demand.
Xue et al. [84]	ECANet SENet	ASVspoof2019	EER = 0.65%	Improves accuracy without large models. Efficient inference.	Limited real-world evaluation. Single benchmark validation.
Xie et al. [85]	CSAM, LCNN, AASIST	Codec fake	EER = 0.616%	Addresses codec-based attacks. Strong robustness to modern deepfakes.	Speech-focused only. Limited noisy environment evaluation.
Tran et al. [37]	Multiconv	ASVspoof2019/ 2021	EER = 0.08%	Excellent multilingual generalization. Robust across diverse datasets.	Requires diverse training data. Computationally expensive.

Note: EER = equal error rate, ML = machine learning, CNN = convolutional neural network, BiLSTM = bidirectional long short-term memory, LCNN = light convolutional neural network, logical access = LA.

Collectively, these studies illustrate a clear evolution in audio deepfake detection research from the use of handcrafted spectral features and CNNs to SSL-based representation learning, attention-driven architectures, ensemble fusion strategies, and generalization-focused training. Currently, the most effective and scalable direction for combating advanced generation techniques is offered by the integration of pretrained speech models with specialized classifiers.

In Table 3, a summary of various research studies focused on the detection of spoofed and AI-generated speech using DL methods is provided, showing the characteristics of these approaches.

6. HYBRID MODEL-BASED AUDIO DEEPFAKE DETECTION

Detection of audio deepfakes using the hybrid model combines two different paradigms, traditional ML techniques and DL, to combine their respective strengths. The goal of hybrid audio deepfake detection systems is not solely focused on either the use of handcrafted acoustic features or end-to-end representation learning; instead, they are trained using multiple levels of feature representation in a heterogeneous ensemble of models that are combined in multiple stages as part of the decision process to make them more robust and to generalize better against increasingly sophisticated spoofing attacks. Hybrid systems can be generally grouped into three categories: (1) the use of ML and DL together; (2) the combination of different DL architectures; and (3) the concatenation of features of different types.

6.1 Integration of machine learning and deep learning

The first hybrid methods combined neural feature learning with traditional classifiers. One representative direction integrated magnitude-spectrum and relative phase-shift features (RPS) with a DNN and a one-class support vector machine (OC-SVM) [86]. The DNN provided two complementary forms of information; its output scores were used for one-class classification, while bottleneck representations were employed to characterize the distribution of genuine speech. While the DNN classifier performed better than one-class classification models when evaluated independently, the combination of scores yielded statistically significant improvements and reduced the EER to below 0.1% for most of the evaluated spoofing attacks [86]. This result suggests that neural and one-class decision mechanisms capture different aspects of the distinction between genuine and manipulated speech. However, the system was still relying on manually designed spectral and phase features.

It has also been demonstrated via comparative ML-DL performance tests that the optimal classifier could depend on the recording environment and dataset creation. The use of combinations of MFCCs, spectral coefficients, zero-crossing rates, and energy features was evaluated using SVM, LSTM, and VGG16 classifiers; the relative performance of the models changed according to recording condition [87]. Especially, VGG16 achieved an accuracy of approximately 93% on the original FoR recordings, while SVM reached approximately 98% on the re-recorded subset. Thus, it can be said that DNNs cannot necessarily outperform classical classifiers under all conditions. The efficiency of both models is conditioned by their feature extraction and inherent inductive bias compatibility with channel distortions and re-recording noise present in the evaluation data.

Hybridization can also occur at the feature transformation stage rather than only at the final decision stage. The integration of non-negative factorization with MFCC representations increased the discriminative quality of feature space before classification with Gaussian Naïve Bayes, Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbors (K-NN). The resulting framework achieved 99.93% accuracy on a publicly available binary deepfake voice database [88]. The results clearly show the significance of increasing the discriminative power of relatively simple classifiers by improving their feature. Nevertheless, such an outstanding result is insufficient for demonstrating robustness across different generators, accents, recording devices, languages, or acoustic environments.

A related development involved multi-stage pipelines that process handcrafted and learned representations in parallel. A two-stage system included waveform and wavelet-based preprocessing steps followed by parallel classification of wavelet and ResNet-50 model trained on Spectro-temporal features [89]. This architecture demonstrated approximately 85.94% accuracy when evaluated on the DEEP-VOICE database. Although this result was lower than those reported by some feature transformation systems, the design demonstrates how handcrafted and learned representations can be processed jointly to capture complementary spoofing evidence.

In conclusion, the combination of ML and DL has progressed from DNN-assisted one-class classification to condition-dependent classifier comparison, feature transformation, and parallel multi-stage pipeline processing [86-89]. These systems demonstrate that complementary decision-making systems enhance detection when each system provides unique data. The disadvantages that they all have in common are dependence on predefined preprocessing, dataset-specific optimization, and manually selected acoustic features. Consequently, their generalizability to unseen generation methods and uncontrolled real-world recordings remains uncertain.

6.2 Fusion of multiple deep learning architectures

A second hybrid direction combines multiple DL architectures to capture different aspects of speech signals. While CNNs are effective for extracting local spectral patterns, recurrent networks, transformers, and attention mechanisms are more suitable for modeling temporal and contextual dependencies. The benefits of the combined architecture were demonstrated in one of the first studies of CNN–RNN architectures that showed how the mixed approach can be beneficial for learning local filters and temporal correlations. Such an architecture based on time-frequency representations was able to achieve the EER of 6.73% on ASVspoof2017, demonstrating effectiveness under previously unseen replay conditions [90]. The later CNN-BiLSTM and CNN-LSTM networks provided additional evidence of the advantages of spatial-temporal fusion. By employing convolutional layers to extract local features and bi-directional or regular recurrent layers to model sequential dependencies, they achieved an accuracy higher than 94% in ASVspoof2019, FoR, and SceneFake challenges [66, 91]. This demonstrates the applicability of sequential modeling to complement the CNN-based spectral analysis. Nevertheless, limited training data and insufficient out-of-domain evaluation make it difficult to determine whether these models are

learning general spoofing features or dataset-specific artifacts.

Transformer-enhanced residual networks (TE-ResNet) represent a further attempt to combine local acoustic processing with global contextual modeling. A Transformer encoder integrated with a ResNet, data augmentation, and logistic-regression-based score fusion achieved EERs of 3.99% on FoR-norm and 5.89% on ASVspoof2019 LA. Even though cross-dataset testing serves as a more reliable measure of generalization than the single-dataset one, the still-present performance difference proves the insufficiency of contextual modeling to address the domain mismatch problem [92].

Compact convolutional transformers are a more lightweight alternative to large transformer architectures. By combining convolutional feature extraction with attention-based modeling of long-range dependencies, one compact architecture obtained 92.13% accuracy, 93.79% precision, and 92.13% recall on spectral representations [93]. Thus, it was demonstrated that the combination of convolution and attention may yield complementary features. However, because the evaluation was limited to a single benchmark, the reported performance does not establish robustness under external or real-world conditions. Hierarchical and multi-stage SSL architectures are another more advanced phase of hybrid architecture development. A two-stage Wav2DF model with a Hierarchical Adaptive Mixture-of-Experts (HA-MOE) module incorporated multi-level spoofing cues and yielded EERs of 0.87% for ASVspoof2021 LA, 0.95% for ASVspoof2021 DF, and 6.83% for in-the-wild [94]. The substantially higher EER under in-the-wild conditions is particularly important because it shows that near-optimal benchmark performance may not transfer directly to realistic acoustic environments. Nevertheless, the architecture demonstrates the potential of multi-level feature fusion and expert routing for handling heterogeneous spoofing cues.

The above-mentioned evidence is thus an indication of development from CNN-RNN fusion to transformers with residual networks, compact convolutional transformers, and hierarchical self-supervised mixture-of-experts systems (SSL MOEs) [66, 90-94]. This type of model improves representation diversity by leveraging spectral, temporal dependencies, contextual relationships, and multi-level information. Disadvantages of such models include the higher cost of computation, complex architecture, reliance on optimization techniques, and low robustness under out-of-distribution or real-world acoustic conditions.

6.3 Multi-feature concatenation and ensemble strategies

These methods employ multi-representation fusion and ensemble strategies to address the limitations of single-view systems by combining different acoustic representations or model predictions. By integrating spectral, cepstral, wavelet, pretrained, or auditory-inspired representations. The underlying rationale is that each representation captures different artifacts; therefore, combining them may improve the system's detection capability. Model-level ensembles combine the decisions of systems trained on different views of the same audio signal.

A multi-spectral framework was proposed that considers short-time Fourier transform (STFT), constant-Q transform (CQT), wavelet transform, and auditory-inspired filter representations with CNN, RNN, and CRNN models [95]. In addition, pretrained embeddings from systems such as Whisper and SpeechBrain have also been included in select model combinations. The resulting ensemble system achieved an EER of 0.03% on ASVspoof2019, illustrating that a combination of different spectral transformations and pretrained representations can provide complementary spoofing information. However, the performance is observed on a single dataset, and executing several feature extractors and classifiers introduces considerable computational and memory requirements.

A two-stage fusion was applied in parallel spectral views before classification, known as feature-level fusion. A model consisting of dual pathways that extract STFT-based and Mel scale-based convolutions, further refined using self-attention before ResNet classification, was proposed [96]. This approach decreased EER by more than 60% compared to its baseline on ASVspoof2021. The approach proves the claim that frequency-based representations may have different information on spoofing. Yet, self-attention, in addition to dual feature extraction, complicates the architecture, while the EER of 8.94% is a sign that there is much space for improvement.

Fusion of cepstral features and temporal features combines classical acoustic features with DL models for sequences. An architecture utilizing the fusion of MFCC and LFCC with VGG16, temporal convolutional networks, and BiLSTM modeling achieved 98.87% accuracy on the in-the-wild database and showed strong performance on the ASVspoof2019 dataset [97]. The architecture benefits from combining local spectral analysis, temporal convolution, and recurrent dependency modeling. Yet, the multilingual capabilities, inference latency, and real-time deployment of the model were not evaluated, limiting conclusions regarding its operational applicability.

More complex hybrid models use the combination of wavelet packet decomposition, temporal features, prosodic information, voice quality features, and MFCC features in parallel LSTM autoencoders and dynamic residual difference units [98]. Evaluations across different datasets were conducted, yielding accuracies from 90% to 97% with an overall AUC of 98%. The use of multiple datasets to validate a model provides more evidence compared to validation through a single benchmark. However, differences in performance across different datasets demonstrate that domain mismatch is still an issue in the proposed architecture, along with complexity and higher costs due to many extracted features and parallel processing.

Lightweight multilingual models attempt to address this problem by providing support for multiple languages without increasing the computational cost. The use of a pruned CNN-Transformer architecture to process Mel, LFCC, and Wav2Vec features for Marathi, Hindi, Tamil, and Telugu voices yielded more than 98% accuracy [99]. The use of pruning is clearly an effort to find the trade-off between performance and efficiency. In conclusion, there is evidence that the use of multi-feature fusion and ensemble techniques allows better detection of audio deepfakes based on the complementary usage of different spectral, temporal, and linguistic features [95-99]. The evolution of the field of study goes from multi-spectral ensembles through dual path feature fusion, cepstral temporal fusion, rich handcrafted and neural representations, and lightweight multilingual models. The main issues in this regard are related to redundant data, computationally expensive algorithms, optimization problems, and poor interpretability. The details of research studies involving hybrid approaches for detecting both spoofed and AI-generated speech are provided in Table 4.

Table 4. Comparative summary of hybrid audio deepfake detection

References	Method	Dataset	Results	Advantage	Limitation
Villalba et al. [86]	DNN One Class SVM	ASVspoof2015	EER < 0.1%	ML-DL fusion outperforms individual models. Better robustness than one-class SVM alone.	Limited generalization to some unseen attacks. Relies on handcrafted features
Hamza et al. [87]	ML + DL	FoR	Accuracy = 98.83%	Hybrid ML-DL improves detection over a single model. Benefits from complementary acoustic features.	Limited evaluation under noise conditions. Dataset-dependent performance.
Gujjar et al. [88]	GNBXtract Net + GNB	Public binary; Deepfake voice	Accuracy = 99.93%	Feature transformation improves classifier performance. Higher accuracy than conventional ML classifiers.	Limited cross-accent generalization. Sensitive to acoustic variability.
Abdelhamid et al. [89]	ANN ResNet50	DEEP-VOICE	Accuracy = 85.94%	Combines interpretable ML with powerful DL. Better feature diversity than single-stage models.	Evaluated on one dataset. Limited evidence of generalization.
Lavrentyeva et al. [90]	CNN + RNN	ASVspoof2017	EER = 6.73%	CNN-RNN captures complementary spectral and temporal cues. More effective than CNN alone.	Evaluated on a single dataset. Limited robustness analysis.
Wani et al. [91]	CNN BiLSTM	ASVspoof2019; FoR	Accuracy = 97.82%	CNN + BiLSTM improves spatio-temporal learning. Better than single-network architectures.	Limited cross-dataset evaluation. Generalization not fully validated.
Bhagat and Borge. [66]	CNN LSTM	SceneFake	Accuracy = 94.8%	Combining CNN with BiLSTM enhances detection accuracy. Exploits complementary feature learning.	Limited dataset diversity. Weak generalization analysis.
Zhang et al. [92]	TE-ResNet	ASVspoof2019; FoR	EER = 5.89% EER = 3.99%	CNN-Transformer improves unseen attack detection. Better contextual modeling than CNNs.	Cross-dataset performance remains limited. Higher computational cost.
Bartusiak and Delp [93]	CCT	ASVspoof2019	Accuracy = 92.13%	Combines CNN with transformer to capture local and global features. More effective than CNN alone.	Evaluated on one benchmark. No cross-dataset validation.
Hao et al. [94]	Wav2vec2 DF-TSL	ASVSpoof2019/ 2021; in-the-wild	EER = 0.103%	Two-stage SSL improves robustness over single-stage models. Strong cross-dataset performance.	High model complexity. Expensive inference.
Pham et al. [95]	CNN, RNN, C-RNN, transfer learning, Audio embedding	ASVspoof2019	Accuracy = 90.0% EER = 0.03%	Ensemble learning improves robustness. Multi-representation fusion outperforms single features.	Evaluated on one benchmark. High computational cost.
Huang and Pun [96]	ResNet	ASVspoof2021	EER = 8.94%	Hybrid spectral features improve discrimination. Better than single feature representations.	Self-attention increases complexity. Computationally demanding.
Momu et al. [97]	VGG16, TCN, BiLSTM	In-the-wild; ASVSpoof2019	Accuracy = 98.87%	Multi-feature fusion improves robustness. Better than individual feature pipelines.	Limited multilingual evaluation. Real-time deployment not validated.
Muruganadham et al. [98]	LSTM-AE-DRDE	CVoice Fake FoR	Accuracy = 97%	Rich feature fusion improves detection accuracy. Better representation than single-feature model.	High ERR on unseen datasets. Complex architecture.
Gaikawad and Ghosh [99]	CNN Transformer	Marathi; Hindi; Tamil; Telugu; ASVspoof2021; in-the-wild	Accuracy = 98.07%	CNN-Transformer improves multilingual detection. Balances accuracy and efficiency.	Requires a pretrained model. Increased computational complexity.

Note: DNN = deep neural network, SVM = support vector machine, EER = equal error rate, ML = machine learning, DL = deep learning, CNN = convolutional neural network, BiLSTM = bidirectional long short-term memory.

7. DISCUSSION

This study was conducted to determine the effectiveness of DL techniques for detecting audio deepfakes and to assess the pace of progress in deepfake detection since DL architectures were developed. In the past, detection was performed using traditional ML algorithms along with handcrafted acoustic features (e.g., MFCC, LFCC, global spectral feature descriptors). While detection methods were successful at detecting deepfakes in controlled laboratory environments, the effectiveness of these types of algorithms would often decrease when working in a real-world setting, which is characterized by background noise, unknown attacks, and cross-dataset analysis. There is now a clear trend among researchers toward developing deep-learning-based detection systems. Current research focuses primarily on CNN, RNN, and transformer-style architectures. CNN architectures have been effective in capturing spatial characteristics, and RNN architectures (e.g., LSTM, Bi-LSTM) have been used to model the temporal characteristics of audio signals. The use of hybrid detection frameworks that incorporate both types of architectures will probably continue to improve deepfake detection accuracy because these types of frameworks allow for simultaneous modeling of spatial and temporal characteristics of sound signals.

Additionally, a major finding from the literature is that

feature fusion methods have become more popular. Many of the best audio processing systems today are developed by combining multiple types of acoustic features, including MFCCs, LFCCs, Mel-spectrograms, and wavelet-based features. Thus, fusing different features allows models to gather complementary information from the audio signal, thereby creating more resistant models to spoofing attempts using different methodologies. Furthermore, attention mechanism(s) and transformer encoder integration methods have also helped to improve model performance by enabling models to concentrate on the most relevant regions within the spoken material. yielding encouraging results.

The benchmark datasets (ASVspoof, FoR, and in-the-wild) have been widely used to evaluate detection algorithms. Many models perform extremely well on a particular dataset, but fail to generalize to unseen attack methods across different recording conditions. As a result, limited diversity and dataset bias continue to pose significant challenges to the development of generalized detection approaches.

A major challenge is the ongoing, fast-paced development of deepfake generation processes. With today's advanced TTS and voice swap (speech synthesis) technology, it is very easy to produce realistic-sounding speech from text, which makes it difficult to distinguish between true and false with existing solutions. Therefore, as generative models (ML-based systems) advance, detection systems will also need to evolve by incorporating stronger features, SSL, and cross-domain training.

In addition, there is growing evidence from a number of studies that models used to detect objects need to be lightweight and efficient. Although DL models with many layers are often used to achieve improved performance, they will not provide acceptable results in real-time environments due to the limitations of the computational power of the computers running the models. For this reason, there is continued effort in researching how to develop efficient and effective models for real-time applications, such as voice authentication systems or embedded hardware systems.

Even with many positive advances within this field, there still exists a shortage of good research, including limited interpretability of models, vulnerability to adversarial attacks, and a lack of diverse, high-quality multilingual datasets representative of real-world audio conditions. In order to address these issues, it will take collaboration between ML, cybersecurity, and digital forensics experts.

Improved robustness and generalization ability of deepfake audio detection systems are also possible through the use of hybrid DL architectures, feature fusion methods, and self-supervised representations. Future research efforts should focus on developing more transparent, adaptable, and dataset-independent detection frameworks that can keep up with the fast-changing landscape of synthetic speech generation technology over time.

8. CONCLUSION

This article provides an overview of three types of audio deepfake technology: generation methods, benchmark datasets, and deepfake detection methods. It also discusses some of the major challenges to using these technologies in real-world settings. Rapid advancements in TTS and VC models have created even more realistic synthetic speech than has been seen in the past few years. This technological progress has both enhanced the risks surrounding identity theft, security of ASV systems, and reliability of digital forensic analysis.

There have been significant advancements in detecting audio deepfakes, especially with the use of DL and SSL techniques. However, the various methods for detecting deepfake audio still cannot keep up with the many advancements in generative audio deepfake creation. The most traditional type of ML will still work well when a well-defined set of engineered features is used to support its operation, particularly in controlled environments where ML is typically applied. On the other hand, more advanced types of ML, such as the use of self-supervised representations and attention-based/hybrid architectures, have been shown to outperform traditional types of ML and generalize better across different scenarios than traditional techniques.

Current detection systems exhibit many limitations regarding adversarial attacks, language diversity, complex real-world acoustic conditions, dataset distinctions, and a lack of discernibility in the detections made. There is an urgent requirement for greater quantities of more diverse and/or more multilingual data, which represent realistic recording conditions and include new emergence techniques for carrying out attacks.

REFERENCES

[1] Altuncu, E., Franqueira, V.N.L., Li, S. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. Frontiers in Big Data, 7: 1400024. https://doi.org/10.3389/fdata.2024.1400024

[2] Mcuba, M., Singh, A., Ikuesan, R.A., Venter, H. (2023). The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science, 219: 211-219. https://doi.org/10.1016/j.procs.2023.01.283

[3] Mohammed, A. (2024). Deepfake detection and mitigation: Securing against AI-generated manipulation. Journal of Computational Innovation, 4(1): 1-25. https://researchworkx.com/index.php/jci/article/view/55.

[4] Kietzmann, J., Lee, L.W., McCarthy, I.P., Kietzmann, T.C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2): 135-146. https://doi.org/10.1016/j.bushor.2019.11.006

[5] Li, X., Chen, P.Y., Wei, W. (2025). Where are we in audio deepfake detection? A systematic analysis over generative and detection models. ACM Transactions, 25(3): 1-19. https://doi.org/10.1145/3736765

[6] Almutairi, Z., Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. Algorithms, 15(5): 155. https://doi.org/10.3390/a15050155

[7] Jia, Y., Zhang, Y., Weiss, R.J., et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 4485-4495.

[8] Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A., Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, 53(4): 3974-4026. https://doi.org/10.1007/s10489-022-

03766-z
[9] Pham, L., Lam, P., Tran, D., et al. (2025). A comprehensive survey with critical analysis for deepfake speech detection. Computer Science Review, 57: 100757. https://doi.org/10.1016/j.cosrev.2025.100757
[10] Khanjani, Z., Watson, G., Janeja, V.P. (2021). How deep are the fakes? Focusing on audio deepfake: A survey. arXiv preprint arXiv:2111.14203. https://doi.org/10.48550/arXiv.2111.14203
[11] Galyashina, E., Nikishin, V. (2021). AI generated fake audio as a new threat to information security: Legal and forensic aspects. In Proceedings of the International Scientific and Practical Conference on Computer and Information Security, Yekaterinburg, Russian Federation, pp. 17-21. https://doi.org/10.5220/0010616700003170
[12] Feng, Z., Chen, J., Zhou, C., et al. (2025). Enkidu: Universal frequential perturbation for real-time audio privacy protection against voice deepfakes. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, pp. 11638-11647. https://doi.org/10.1145/3746027.3755629
[13] Salih, A.O.M., Emam, A.H.M., Ahmed, A.B.G.E., Khalifa, M., Suliman, A., Babiker, N.B.M. (2025). Deepfake audio detection in voice authentication: A spectral and CNN-based comprehensive review. Engineering, Technology & Applied Science Research, 15(6): 29824-29832. https://doi.org/10.48084/etasr.13400
[14] Zhang, B., Cui, H., Nguyen, V., Whitty, M. (2025). Audio deepfake detection: What has been achieved and what lies ahead. Sensors, 25(7): 1989. https://doi.org/10.3390/s25071989
[15] Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., Parkinson, S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. IEEE Access, 11: 144497-144529. https://doi.org/10.1109/ACCESS.2023.3344653
[16] Verdoliva, L. (2020). Media forensics and deep fakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5): 910-932. https://doi.org/10.1109/JSTSP.2020.3002101
[17] Chesney, B., Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107: 1753. https://doi.org/10.2139/ssrn.3213954
[18] Mai, K.T., Bray, S.D., Davies, T., Griffin, L.D. (2023). Warning: Humans cannot reliably detect speech deepfakes. Plos One, 18(8): e0285333. https://doi.org/10.1371/journal.pone.0285333
[19] Kaur, N., Dixit, A., Kingra, S. (2025). A deep learning fusion model leveraging spectral features for audio deepfake detection. International Journal of Advanced Networking and Applications, 16(5): 6551-6560. https://doi.org/10.35444/IJANA.2025.16503
[20] Riaz, S., Tariq, A., Arif, E., et al. (2025). The advanced AI techniques for deepfake audio detection. Journal of Computing & Biomedical Informatics, 9(2). https://jcbi.org/index.php/Main/article/view/1059.
[21] Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C.Y., Zhao, Y. (2023). Audio deepfake detection: A survey. arXiv preprint arXiv:2308.14970. https://doi.org/10.48550/arXiv.2308.14970
[22] Babaei, R., Cheng, S., Duan, R., Zhao, S. (2025). Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. Journal of Sensor and Actuator Networks, 14(1): 17. https://doi.org/10.3390/jsan14010017
[23] Tan, X., Qin, T., Soong, F., Liu, T.Y. (2021). A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561. https://doi.org/10.48550/arXiv.2106.15561
[24] Taylor, P. (2009). Text-to-Speech Synthesis. Cambridge University Press.
[25] Wang, Y., Skerry-Ryan, R.J., Stanton, D., et al. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135. https://doi.org/10.48550/arXiv.1703.10135
[26] Zen, H. (2015). Acoustic modeling in statistical parametric speech synthesis: From HMM to LSTM-RNN. Proceedings of MLSLP, 15. https://research.google.com/pubs/archive/43893.pdf.
[27] Goyal, A.K., Kumari, A., Raj, A., Gautam, S., Ishita, Raghuvanshi, D. (2024). TalkifyPy: The pythonic voice assistant. In 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET), Ghaziabad, India, pp. 1-5. https://doi.org/10.1109/ACET61898.2024.10730081
[28] Orynbay, L., Razakhova, B., Peer, P., Meden, B., Emeršič, Ž. (2024). Recent advances in synthesis and interaction of speech, text, and vision. Electronics, 13(9): 1726. https://doi.org/10.3390/electronics13091726
[29] Barrault, L., Chung, Y.A., Meglioli, M.C., et al. (2023). Seamlessm4t: Massively multilingual & multimodal machine translation. arXiv preprint arXiv:2308.11596. https://arxiv.org/abs/2308.11596
[30] Sun, X., Wang, Z. (2024). Intelligent English automatic translation system based on TTS technology. In 2024 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS), Bristol, United Kingdom, pp. 687-692. https://doi.org/10.1109/AIARS63200.2024.00130
[31] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. Speech Communication, 66: 130-153. https://doi.org/10.1016/j.specom.2014.10.005
[32] AlAli, A., Theodorakopoulos, G. (2023). An RFP dataset for real, fake, and partially fake audio detection. In International Conference on Cyber Security, Privacy in Communication Networks, pp. 1-15. https://doi.org/10.1007/978-981-97-3973-8_1
[33] Felps, D., Bortfeld, H., Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. Speech Communication, 51(10): 920-932. https://doi.org/10.1016/j.specom.2008.11.004
[34] Sisman, B., Yamagishi, J., King, S., Li, H. (2021). An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 132-157. https://doi.org/10.1109/TASLP.2020.3038524
[35] Walczyna, T., Piotrowski, Z. (2023). Overview of voice conversion methods based on deep learning. Applied Sciences, 13(5): 3100. https://doi.org/10.3390/app13053100
[36] Kobayashi, K., Toda, T. (2018). Electrolaryngeal speech enhancement with statistical voice conversion based on

CLDNN. In 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, pp. 2115-2119. https://doi.org/10.23919/EUSIPCO.2018.8553154

[37] Kinnunen, T., Wu, Z.Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, pp. 4401-4404. https://doi.org/10.1109/ICASSP.2012.6288895

[38] Tran, H.M., Lolive, D., Sini, A., Delhay, A., Marteau, P.F., Guennec, D. (2025). Multi-level SSL feature gating for audio deepfake detection. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, pp. 11766-11775. https://doi.org/10.1145/3746027.3754568

[39] Liu, S., Wu, H., Lee, H.Y., Meng, H. (2019). Adversarial attacks on spoofing countermeasures of automatic speaker verification. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, pp. 312-319. https://doi.org/10.1109/ASRU46091.2019.9003763

[40] Liu, X., Sahidullah, M., Lee, K.A., Kinnunen, T. (2024). Generalizing speaker verification for spoof awareness in the embedding space. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 1261-1273. https://doi.org/10.1109/TASLP.2024.3358056

[41] Todisco, M., Panariello, M., Wang, X., Delgado, H., Lee, K.A., Evans, N. (2024). Malacopula: Adversarial automatic speaker verification attacks using a neural-based generalised Hammerstein model. arXiv preprint arXiv:2408.09300. https://doi.org/10.48550/arXiv.2408.09300

[42] Panariello, M., Ge, W., Tak, H., Todisco, M., Evans, N. (2023). Malafide: A novel adversarial convolutive noise attack against deepfake and spoofing detection systems. arXiv preprint arXiv:2306.07655. https://doi.org/10.48550/arXiv.2306.07655

[43] Rabhi, M., Bakiras, S., Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Systems with Applications, 250: 123941. https://doi.org/10.1016/j.eswa.2024.123941

[44] Wu, Z., Yamagishi, J., Kinnunen, T., et al. (2017). ASVspoof: The automatic speaker verification spoofing and countermeasures challenge. IEEE Journal of Selected Topics in Signal Processing, 11(4): 588-604. https://doi.org/10.1109/JSTSP.2017.2671435

[45] Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K.A., Yamagishi, J. (2018). ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements. In the Speaker and Language Recognition Workshop, pp. 296-303.

[46] Wang, X., Yamagishi, J., Todisco, M., et al. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language, 64: 101114. https://doi.org/10.1016/j.csl.2020.101114

[47] Yamagishi, J., Wang, X., Todisco, M., et al. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. arXiv preprint arXiv:2109.00537. https://doi.org/10.48550/arXiv.2109.00537

[48] Wang, X., Delgado, H., Tak, H., et al. (2024). ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. arXiv preprint arXiv:2408.08739. https://doi.org/10.48550/arXiv.2408.08739

[49] Reimao, R., Tzerpos, V. (2019). FoR: A dataset for synthetic speech detection. In 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, pp. 1-10. https://doi.org/10.1109/SPED.2019.8906599

[50] Frank, J., Schönherr, L. (2021). Wavefake: A data set to facilitate audio deepfake detection. arXiv preprint arXiv:2111.02813. https://doi.org/10.48550/arXiv.2111.02813

[51] Yi, J., Fu, R., Tao, J., et al. (2022). Add 2022: The first audio deep synthesis detection challenge. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 9216-9220. https://doi.org/10.48550/arXiv.2202.08433

[52] Müller, N.M., Czempin, P., Dieckmann, F., Froghyar, A., Böttinger, K. (2022). Does audio deepfake detection generalize? arXiv preprint arXiv:2203.16263. https://doi.org/10.48550/arXiv.2203.16263

[53] Müller, N.M., Kawa, P., Choong, W.H., et al. (2024). Mlaad: The multi-language audio anti-spoofing dataset. In 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, pp. 1-7. https://doi.org/10.1109/IJCNN60899.2024.10650962

[54] Zhang, Y., Zang, Y., Shi, J., et al. (2024). Svdd challenge 2024: A singing voice deepfake detection challenge evaluation plan. arXiv preprint arXiv:2405.05244. https://doi.org/10.48550/arXiv.2405.05244

[55] Wu, Z., Chng, E.S., Li, H. (2012). Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In 13th Annual Conference of the International Speech Communication Association.

[56] Hanilçi, C., Kinnunen, T., Sahidullah, M., Sizov, A. (2015). Classifiers for synthetic speech detection: A comparison. In Interspeech 2015, pp. 2057-2061. https://doi.org/10.21437/Interspeech.2015-466

[57] Liu, Y., Tian, Y., He, L., Liu, J., Johnson, M.T. (2015). Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing. In Interspeech 2015, pp. 2082-2086. https://doi.org/10.21437/Interspeech.2015-471

[58] Rodríguez-Ortega, Y., Ballesteros, D.M., Renza, D. (2020). A machine learning model to detect fake voice. In Applied Informatics, Communications in Computer and Information Science, 1277: 3-13. https://doi.org/10.1007/978-3-030-61702-8-1

[59] Singh, A.K., Singh, P. (2021). Detection of AI-synthesized speech using cepstral and bispectral statistics. In 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), Tokyo, Japan, pp. 412-417. https://doi.org/10.1109/MIPR51284.2021.00076

[60] Borrelli, C., Bestagini, P., Antonacci, F., Sarti, A., Tubaro, S. (2021). Synthetic speech detection through short-term and long-term prediction traces. EURASIP Journal on Information Security, 2021(1): 2. https://doi.org/10.1186/s13635-021-00116-3

[61] Iqbal, F., Abbasi, A., Javed, A.R., Jalil, Z., Al-Karaki, J. (2022). Deepfake audio detection via feature engineering and machine learning. In CIKM workshops, 3318: 1-12. https://ceur-ws.org/Vol-3318/paper4.pdf?

[62] Alegre, F., Amehraye, A., Evans, N. (2013). A one-class

classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Arlington, VA, USA, pp. 1-8. https://doi.org/10.1109/BTAS.2013.6712706

[63] Bisogni, C., Loia, V., Nappi, M., Pero, C. (2024). Acoustic features analysis for explainable machine learning-based audio spoofing detection. Computer Vision and Image Understanding, 249: 104145. https://doi.org/10.1016/j.cviu.2024.104145

[64] Bohara, R., Bairwa, A.K. (2025). Detecting deepfake audio using spectrogram-based machine learning approaches. IEEE Access, 13: 149478-149489. https://doi.org/10.1109/ACCESS.2025.3602531

[65] Li, M., Ahmadiadli, Y., Zhang, X.P. (2025). A survey on speech deepfake detection. ACM Computing Surveys, 57(7): 1-38. https://doi.org/10.1145/3714458

[66] Bhagat, T., Borge, N. (2025). Towards secure audio: Deepfake detection with CNN and LSTM networks. International Journal of Scientific Research in Engineering and Management, 9(4): 1-9. https://doi.org/10.55041/IJSREM44082

[67] Alzantot, M., Wang, Z., Srivastava, M.B. (2019). Deep residual neural networks for audio spoofing detection. In Interspeech 2019, pp. 1078-1082. https://doi.org/10.21437/Interspeech.2019-3174

[68] Wu, Z., Das, R.K., Yang, J., Li, H. (2020). Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. arXiv preprint arXiv:2009.09637. https://doi.org/10.48550/arXiv.2009.09637

[69] Rahul, T.P., Aravind, P.R., Ranjith, C., Nechiyil, U., Paramparambath, N. (2020). Audio spoofing verification using deep convolutional neural networks by transfer learning. arXiv preprint arXiv:2008.03464. https://doi.org/10.48550/arXiv.2008.03464

[70] Bartusiak, E.R., Delp, E.J. (2022). Frequency domain-based detection of generated audio. arXiv preprint arXiv:2205.01806. https://doi.org/10.48550/arXiv.2205.01806

[71] Shin, H.S., Heo, J., Kim, J.H., Lim, C.Y., Kim, W., Yu, H.J. (2024). Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, pp. 10581-10585. https://doi.org/10.1109/ICASSP48485.2024.10448453

[72] Rosello, E., Gomez-Alanis, A., Gomez, A.M., Peinado, A. (2023). A conformer-based classifier for variable-length utterance processing in anti-spoofing. In Interspeech 2023, pp. 5281-5285. https://doi.org/10.21437/Interspeech.2023-1820

[73] Chen, Y., Yi, J., Xue, J., et al. (2024). Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. arXiv preprint arXiv:2406.06086. https://doi.org/10.48550/arXiv.2406.06086

[74] Tak, H., Todisco, M., Wang, X., Jung, J.W., Yamagishi, J., Evans, N. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. arXiv preprint arXiv:2202.12233. https://doi.org/10.48550/arXiv.2202.12233

[75] Wang, X., Yamagishi, J. (2021). Investigating self-supervised front ends for speech spoofing countermeasures. arXiv preprint arXiv:2111.07725. https://doi.org/10.48550/arXiv.2111.07725

[76] Wang, C., Yi, J., Tao, J., et al. (2022). Fully automated end-to-end fake audio detection. In Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal, pp. 27-33. https://doi.org/10.1145/3552466.3556530

[77] Combei, D., Stan, A., Oneata, D., Cucu, H. (2024). WavLM model ensemble for audio deepfake detection. arXiv preprint arXiv:2408.07414. https://doi.org/10.48550/arXiv.2408.07414

[78] Xie, Y., Zhang, Z., Yang, Y. (2021). Siamese network with wav2vec feature for spoofing speech detection. In Interspeech 2021, pp. 4269-4273. https://doi.org/10.21437/Interspeech.2021-847

[79] Conti, E., Salvi, D., Borrelli, C., et al. (2022). Deepfake speech detection through emotion recognition: A semantic approach. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 8962-8966. https://doi.org/10.1109/ICASSP43922.2022.9747186

[80] Krishnan, K.S., Krishnan, K.S. (2023). MFAAN: Unveiling audio deepfakes with a multi-feature authenticity network. In 2023 9th International Conference on Signal Processing and Communication (ICSC), NOIDA, India, pp. 585-590. https://doi.org/10.1109/ICSC60394.2023.10441405

[81] Kanwal, T., Mahum, R., AlSalman, A.M., Sharaf, M., Hassan, H. (2024). Fake speech detection using VGGish with attention block. Journal of Audio, Speech, and Music Processing, 2024(1): 35. https://doi.org/10.1186/s13636-024-00348-4

[82] Feng, Y. (2024). Audios don't lie: Multi-frequency channel attention mechanism for audio deepfake detection. arXiv preprint arXiv:2412.09467. https://doi.org/10.48550/arXiv.2412.09467

[83] Yang, Y., Qin, H., Zhou, H., et al. (2024). A robust audio deepfake detection system via multi-view feature. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, pp. 13131-13135. https://doi.org/10.1109/ICASSP48485.2024.10446560

[84] Xue, J., Fan, C., Yi, J., et al. (2023). Learning from yourself: A self-distillation method for fake speech detection. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10096837

[85] Xie, Y., Lu, Y., Fu, R., et al. (2025). The codecfake dataset and countermeasures for the universally detection of deepfake audio. IEEE Transactions on Audio, Speech and Language Processing, 33: 386-400. https://doi.org/10.1109/TASLPRO.2025.3525966

[86] Villalba, J., Miguel, A., Ortega, A., Lleida, E. (2015). Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In Proceedings Interspeech 2015, Dresden, Germany, pp. 2067-2071. https://doi.org/10.21437/Interspeech.2015-468

[87] Hamza, A., Javed, A.R.R., Iqbal, F., et al. (2022). Deepfake audio detection via MFCC features using machine learning. IEEE Access, 10: 134018-134028. https://doi.org/10.1109/ACCESS.2022.231480

[88] Gujjar, M.U.T., Munir, K., Amjad, M., Rehman, A.U.,

Bermak, A. (2024). Unmasking the fake: Machine learning approach for deepfake voice detection. IEEE Access, 12: 197442-197453. https://doi.org/10.1109/ACCESS.2024.3521026

[89] Abdelhamid, R.M., Naiem, S., Nasr, M.M., Moussa, F.A. (2025). Deepfake audio detection using feature-based and deep learning approaches: ANN vs ResNet50. International Journal of Advanced Computer Science and Applications, 16(6). https://doi.org/10.14569/ijacsa.2025.0160631

[90] Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. In Interspeech 2017, pp. 82-86. https://doi.org/10.21437/Interspeech.2017-360

[91] Wani, T.M., Qadri, S.A.A., Comminiello, D., Amerini, I. (2024). 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, pp. 271-276. https://doi.org/10.1145/3658664.3659647

[92] Zhang, Z., Yi, X., Zhao, X. (2021). Fake speech detection using residual network with transformer encoder. In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, Virtual Event, Belgium, pp. 13-22. https://doi.org/10.1145/3437880.3460408

[93] Bartusiak, E.R., Delp, E.J. (2021). Synthesized speech detection using convolutional transformer-based spectrogram analysis. In 2021 55th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, pp. 1426-1430. https://doi.org/10.1109/IEEECONF53345.2021.9723142

[94] Hao, Y., Chen, Y., Xu, M., et al. (2025). Wav2df-tsl: Two-stage learning with efficient pre-training and hierarchical experts fusion for robust audio deepfake detection. In 2025 International Joint Conference on Neural Networks (IJCNN), Rome, Italy, pp. 1-8. https://doi.org/10.1109/IJCNN64981.2025.11228261

[95] Pham, L., Lam, P., Nguyen, T., Nguyen, H., Schindler, A. (2024). Deepfake audio detection using spectrogram-based feature and ensemble of DL models. In 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2), Erlangen, Germany, pp. 1-5. https://doi.org/10.1109/IS262782.2024.10704095

[96] Huang, L., Pun, C.M. (2024). Self-attention and hybrid features for replay and deep-fake audio detection. arXiv preprint arXiv:2401.05614. https://doi.org/10.48550/arXiv.2401.05614

[97] Momu, S.A., Siddiqui, R.R., Shanto, S.S., Ahmed, Z. (2024). A comprehensive approach to deepfake audio detection: Using feature fusion and deep learning. In 2024 27th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, pp. 351-356. https://doi.org/10.1109/ICCIT64611.2024.11022599

[98] Muruganadham, P., Thangasamy, G.R., Jayaraman, S., Dharmarajan, R. (2025). LSTM autoencoder based parallel architecture for deepfake audio detection with dynamic residual encoding and feature fusion. Scientific Reports, 15(1): 23514. https://doi.org/10.1038/s41598-025-08198-6

[99] Gaikawad, M., Ghosh, S. (2025). A robust and lightweight CNN-transformer model for audio deepfake detection in Indian languages. In 2025 7th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, pp. 382-387. https://doi.org/10.1109/ISPCC66872.2025.11039572