Deepfake Audio Synthesis and Detection: A Review of Recent Advances and Open Challenges

Afifa Yasin, Uswah Fatima, Muhammad Zaid, Asjad Amin
Faculty of Engineering and Technology, The Islamia University of Bahawalpur, Punjab, Pakistan

***Abstract*—There has been tremendous growth in the application of Deep Learning algorithms in recent years, and this has brought about significant advancements in voice synthesis, allowing the possibility of lifelike 'deepfake audio' to be generated with models such as Transformers, GANs, and hybrids. These developments have enriched applications in terms of customization and ease of use but simultaneously bring about high risks associated with cybersecurity, social engineering, and misinformation with respect to 'deepfake audio' because of its possibility due to such advancements. This article review has considered 50 recent studies on 'deepfake voice' creation and detection to discuss its full application in this paper, developed in two stages. Stage 1 introduces synthesis techniques such as 'prosody with attention,' 'diffusion models for diversity,' and 'GAN-based voice synthesis' to describe recent advancements in 'multilingual, zero-shot, and expressive capabilities.Phase 2 is focused on detection techniques that address spectral, temporal, and multimodal artifacts in addition to traditional methods for robustness in low-resource environments. These techniques include Transformer encoders, CNN hybrids, and self-supervised frameworks. Despite significant advancements, challenges with adversarial attacks, multilingual deepfakes, real-world noise, and generalization to unknown models still exist. In order to protect voice-based technologies in an AI-driven era, the review identifies future directions, that are centered around multimodal integration, continuous learning, and interpretable forensics.**

***Index Terms*—Deepfake audio, voice cloning, zero-shot TTS, voice conversion, audio forensics, anti-forensic detection, self-supervised learning, neural codec models**

1. INTRODUCTION

One of the biggest risks to digital authenticity in the present informational ecosystem is the quick spread of deepfake audio technology. Malicious actors can now synthesize or manipulate human voices with unmatched realism by utilizing sophisticated deep learning algorithms. This makes way for a variety of fraudulent applications, such as financial fraud, cyber extortion, political disinformation, and celebrity impersonation. Forged audio clips, for example, have been used to mimic world leaders in fake speeches, impact public confidence during elections, or trick people into sending money under false pretenses.Large datasets and generative models have come together to produce this development, which allows systems to imitate the accents and emotional variations besides timbre and intonation with minimal input data.The universal nature of social media platforms and communication tools raises ethical, legal, and social issues about the decreasing verifiable truth in spoken content due to the increased potential for widespread misuse. Initial alerts about such threats in significant works on voice modification emphasized the need for robust countermeasures to preserve the integrity of audio as evidence [1], [2].

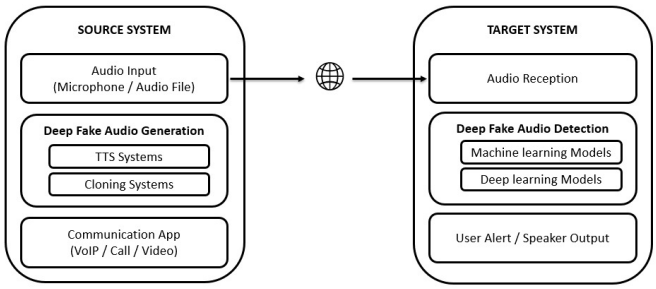

Fig. 1. Process of synthetic based deepfake [3]

Synthesis techniques have progressed from basic recurrent networks like Tacotron [4] to sophisticated GAN-based vocoders like HiFiGAN [5] and diffusion models that produce almost identical results.These advancements enable zero-shot or few-shot voice cloning, as demonstrated by large-scale systems such as VALL-E [6], which can generate speech in unseen voices using brief samples. Methods for detecting subtle artefacts have shifted to multimodal and self-supervised frameworks, such as Whisper-based encoders for cross-lingual robustness [7] and hybrid ensembles that combine CNNs with attention mechanisms for better generalization [8]. However, real-world deployment reveals persistent vulnerabilities, such as vulnerability to adversarial perturbations [9] and performance losses under compression or noise [10].This discrepancy highlights the limitations of current datasets, particularly for low-resource scenarios [11], which often lack diversity in environmental conditions, languages, and accents.

This review paper fills in these gaps by providing an extensive classification of deepfake voice synthesis models, detection methods, and classical machine learning alternatives. It does this by integrating information from over 50 recent studies to present an all-encompassing picture of the field. Our work highlights audio-specific challenges, such as multilingual manipulation and forensic interpretability, while suggesting future directions for robust, real-time systems, in contrast to previous reviews that narrowly focus on visual deepfakes or general AI ethics [12]. We hope to help researchers, policymakers, and practitioners create more potent defenses by classifying the literature into synthesis phases, detection paradigms, and emerging trends.The remainder of the paper is structured as follows: Section 2 details voice synthesis

models,explores deep and classical machine learning-based detection and related surveys, Section 3 discusses challenges and future directions, and Section 4 concludes with key takeaways.

2. RELATED WORK

Phase 1, which focuses on deepfake voice synthesis, and Phase 2, which focuses on deepfake voice detection, are the two main phases of this section's literature review. Based on the reviewed papers, each phase is arranged into the appropriate model categories and subcategories.

2.1 DEEPFAKE VOICE SYNTHESIS

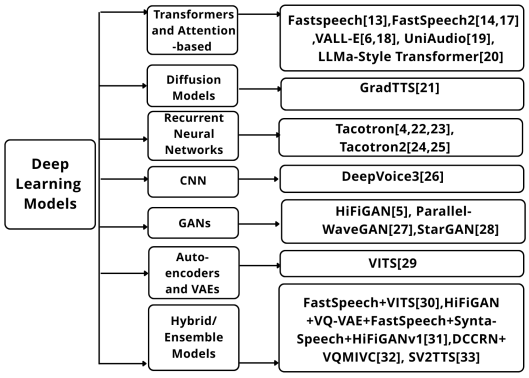

Fig. 2. Phase:1 Deepfake Voice Synthesis

2.1.1 Deep Learning Models

1) Transformers and Attention-Based:

Modern text-to-speech (TTS) and voice cloning systems have made significant strides ,thanks to transformer-based architectures. Many expressive TTS systems rely on FastSpeech and FastSpeech 2, which provide duration, pitch, and energy predictors for extremely controllable prosody [13]. Personalized lightweight TTS models use structured pruning to increase efficiency without sacrificing voice quality [14].Further research improved the controllability of accented TTS [15], multi-speaker expressive synthesis [16],and multimodal adversarial training for zero-shot voice cloning [17]. Neural codec language models like VALL-E and VALL-E 2 use discrete codec tokens rather than mel-spectrograms for high-fidelity zero-shot synthesis [18]; multilingual voice transformation is made possible by cross-lingual neural codec modeling [6]. Large-scale models like VoiceTuner emphasize self-supervised pretraining for efficient fine-tuning [19], whereas SyncSpeech presents low-latency TTS aligned with LLM-style architectures using dual-stream temporal Transformer masking [20]. When taken as a whole, these models demonstrate how attention-driven architectures dominate the production of expressive, multilingual, and controllable synthetic speech.

2) Diffusion Models:

Probabilistic speech generation through sequential refinement of noisy latent representations has been investigated in diffusion-based TTS research. GradTTS uses this reverse-diffusion mechanism to produce high-quality speech from untranscribed data, which makes it very useful for adaptation with limited resources [21]. Diffusion's stochastic properties allow for robust voice reconstruction across a variety of speaking patterns and more natural prosody modeling.

3) Recurrent Neural Networks:

Because of its strong performance in expressive speech modeling and its simple encoder-decoder structure, Tacotron and its variations continue to be influential. Tacotron systems that are multilingual have proven to be flexible in environments with limited resources [22], whereas advanced speaker embeddings in zero-shot multi-speaker TTS highlights the importance of embedding-based speaker identity control [23]. Prosodic clustering techniques improve phoneme-level expressiveness [4], and Tacotron 2 innovations in few-shot cloning and meta-learning frameworks (META-VOICE) further facilitate quick style transfer and speaker adaptation [24], [25].

4) CNN-Based Models:

CNN-based TTS frameworks such as DeepVoice3 utilize convolutional mechanisms to process data in parallel, allowing for large multi-speaker capacity and real-time synthesis [26]. Despite their constraints in simulating long-term dependencies, these models effectively help in robust and scalable voice generation.

5) GAN-Based Models:

Adversarial audio synthesis is still important for amplifying realism and reducing noise artifacts. Systems based on HiFi-GAN introduce noise-robust voice conversion capabilities [5] , while ParallelWaveGAN supports unsupervised cross-domain singing voice conversion [27]. StarGAN-based architectures demonstrate the adaptability of GANs in cross-speaker scenarios by enabling speaker-agnostic one-shot voice conversion through domain translation [28].

6) Autoencoders and Variational Models:

Autoencoder-driven TTS systems such as VITS provide end-to-end probabilistic modeling combining flows and variational inference [29]. Recent work introduced dual-style feature modulation to support zero-shot style adaptation, enabling highly expressive speaker similarity even with limited samples.

7) Hybrid Models:

Hybrid deepfake synthesis integrates the strengths of multiple architectures. Systems combining FastSpeech and VITS enable spontaneous TTS style modeling [30]; others fuse HiFiGAN, VQ-VAE, and FastSpeech to create noise-robust voice conversion [31]. Real-time cloning frameworks have adopted multiple modules such as DCCRN and VQMIVC for enhanced perceptual quality [32]. Large-scale pipelines such as Mega-TTS demonstrate scalable zero-shot TTS via intrinsic inductive biases [33].

These trends are contextualized by a survey on voice cloning systems, which also highlights important constraints such as prosody control, emotional variability, and multilingual performance gaps [34].

2.2 DEEPFAKE VOICE DETECTION

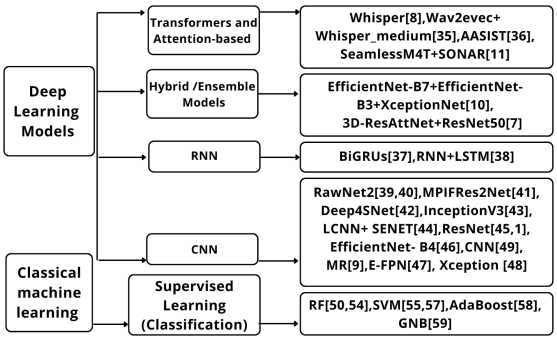

Fig. 3. Phase:2 Deepfake Voice Detection

2.2.1 Deep Learning Models

1)Transformers and Attention-based Models:

Transformer-based detectors use trustworthy, pre-trained encoders to identify small manipulation cues. Singing voice deepfake detection has been successfully accomplished by Whisper-based encodings [7], while Wav2Vec in conjunction with Whisper-medium enhances generalization across various datasets [35]. The significance of voiced versus unvoiced segmentation in identifying synthesis artifacts is highlighted by AASIST-based models [36]. SeamlessM4T and SONAR multimodal frameworks address manipulation in low-resource languages by extending the detection into multilingual and hate-speech scenarios [11].

2)Hybrid/Ensemble Models:

Several works incorporate CNNs, attention modules, and residual structures to perform spectral analysis. EfficientNet, XceptionNet hybrids provide robustness against adversarial perturbations [9], while attention-based 3D-ResAttNet and ResNet50 support audio–visual deepfake analysis. A paper Proposes an ensemble multimodal deepfake detection method to identify auditory and facial manipulations by exploiting correspondence between audio-visual modalities; utilizes unimodal (audio and video) and cross-modal learning networks with an attention-based deep convolutional neural network [8].

3) Recurrent Neural Networks:

BiGRU-based systems utilize cross-modal attention for fine-grained temporal inconsistencies. It Propose a novel multimodal attention framework for robust audio-visual deepfake detection and localization, leveraging contextual information and addressing modality gaps by Introducing Multi-Modal Multi-Sequence Bi-Modal Attention (MMMS-BA), an RNN-based framework [37]. RNN+LSTM architectures tailored for Urdu speech present the importance of linguistic diversity in detection [38].

4) CNN-Based Fake Audio Detection:

Many models remain variants of CNNs due to their strong representation of spectro-temporal artifacts. The use of raw features through HuBERT embeddings strengthens robustness against codec distortions in RawNet2 detectors [39], [40].MPIF-Res2Net models emphasize frequency-mix distillation [41], Deep4SNet emphasizes adversarial resistance [42], and InceptionV3 pursues defensive strategies [43]. Permutation entropy is introduced by LCNN+SENET models to regularize the feature space [44]. Other ResNet-based hybrids include works leveraging spectral enhancement and adversarial training, strengthening the importance of spectral analysis [45], [46]. Noise-reliant detectors such as MR and focused architectures on the presence of compression artifacts such as E-FPN further improve robustness against compression and quality variations [10], [47]. Few-shot contrastive approaches leveraging Xception alleviate issues caused by limited datasets [48]. Artifact simulation-based methods further improve diversity in training data [1]. CNN and Ensemble model has been used with random forest for auditory analysis, GAN-based discriminator, audio-visual spectral analysis [49] .

2.2.2 Classical Machine Learning Models

Deepfake audio detection still heavily relies on traditional machine learning techniques, especially in settings with limited resources or when interpretability is more important than raw performance. These techniques usually use conventional classifiers in conjunction with manually created acoustic features, most frequently Mel-Frequency Cepstral Coefficients (MFCCs).

1) Random Forest-based Approaches:

Random Forest classifiers have been used in conjunction with transfer learning–based feature extraction to detect fake audio scenes, especially in constrained or low-data settings [50]. By extracting MFCC features from audio samples and feeding them into a Random Forest model, there has been proposed a system that achieves reliable detection by learning spectral patterns that differentiate real speech from artificial manipulations [51]. Bohara and Bairwa investigated spectrogram-based representations converted into machine learning-ready features, using Random Forest among ensemble techniques to identify deepfake audio and emphasizing its efficiency in capturing time-frequency disparities [52]. A multi-algorithmic approach demonstrated better generalization across various deepfake generation techniques by integrating Random Forest into a larger ensemble that also takes multimodal cues (audio and video) into account [53]. In order to preserve audio integrity through binary classification, Random Forest has been used on extracted features to precisely unmask artificially generated voices in controlled environments [54].

2) Support Vector Machine-based Approaches:

An SVM classifier trained on MFCC features taken from the FoR dataset was presented, that has highlighted SVM's superiority for separating complex spectral patterns in deepfake audio and reporting accuracy up to 97.28% [55]. SVM outperformed other ML models (including Gradient Boosting) on short audio segments when paired with MFCC features, according to extensive experiments on ASVspoof and FoR datasets [56]. "AudioVeritas," is a lightweight SVM-based model that incorporated multiple acoustic features (MFCCs,

chroma, spectral properties), has achieved strong classification performance suitable for practical deployment [57].

3) AdaBoost-Based Approach:

There has been introduced a new biologically inspired feature called speech pause patterns. It has been achieved by training a *AdaBoost* ensemble by looking at pause duration, frequency, and distribution—features affected by human respiration and cognition—and demonstrated promising discrimination between real and artificial voices, especially from platforms like Descript and ElevenLabs [58].

4)Gaussian Naive Bayes-Based Approach:

There has been employed *Gaussian Naive Bayes* on acoustic features to detect deepfake voices, employing probabilistic modeling for efficient classification in scenarios with limited computational resources [59].

Adversarial attacks, multimodal deepfake detection, and audio deepfake detection summarize systemic issues such as adversarial susceptibility, dataset bias, lack of generalization, and the quickly changing field of synthetic audio technologies [2], [60], [12].

3. CHALLENGES AND FUTURE DIRECTIONS

Despite the fact that the methods of synthesis and detection are not yet finished and continue to evolve quite swiftly, there are various fundamental challenges that remain. It has been shown that the challenge of generalization across the domain remains, where the detectors tend to generally do well on known datasets but experience strong drops when confronted with unfamiliar generative models or mismatch conditions. Other key challenges are robustness to adversarial samples; where even complex hybrid baselines, which consist of CNNs, attention mechanisms, and residual networks, can be affected severely even with small perturbations, and noise-trained methods such as MR and artifacts-aware architectures such as E-FPN highlight the challenges of quality variability and codec and real-world audio distortions introduced by noise, compression, or transmission pathways.

Another factor that has increased the complexity is the availability of multilingual and cross-linguistic deepfakes. In low-resource languages, deepfakes are more complex because of limited training data and intricate language patterns. In artifact simulation and few-shot contrastive approaches, the concern is that current deepfake algorithms are becoming better at eliminating observable artifacts, causing spectro-temporal features to become less useful. Biased datasets and an overall lack of diversity in resources also harm other more established areas, making Random Forest classification systems less effective in forensic analysis. In order to protect against an increase in synthetic audio threats and risks, benchmarking and noise-robust training, as well as an overall ethical framework, are going to be necessary.

4. CONCLUSION

From the early days of autoregressive models and GAN networks to extremely realistic evolving models like the state-of-the-art transformer networks today, these models still only scratch the surface. the fact that existing models are typically very strong at known synthesis techniques but weak on models that have not yet been seen. To close the upcoming gap, much more than small architectural advances will be required. In the upcoming years, it will be necessary to focus on generalizability across tasks and domains, robustness against adversaries with fraudulent goals, and the complexity constraints under practical deployment conditions. This will require a variety of sets of innovative benchmarks, noise-robust training strategies, and open source evaluation structures. It is also crucial to design traceable forensics solutions that can not only identify manipulation of any kind but also produce evidence-based traces that are pertinent to the related legal proceedings. In order to effectively lessen the downstream effects on society for the use of the emerging audio deepfakes, it is ultimately necessary for the emerging solutions to the problem at hand to be proactively governed through multifaceted initiatives from media education and collaborative research and development schemes involving the community at all levels with the help of related technology organizations and relevant policy makers.

REFERENCES

[1] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.

[2] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE access*, 9:155161–155196, 2021.

[3] Abhishek Dixit, Nirmal Kaur, and Staffy Kingra. Review of audio deepfake detection techniques: Issues and prospects. *Expert Systems*, 40, 04 2023.

[4] Anonymous. Prosodic clustering for phoneme-level prosody control in end-to-end speech synthesis. *IEEE ICASSP*, 2021. DOI: 10.1109/ICASSP39728.2021.9413604.

[5] Anonymous. Preserving background sound in noise-robust voice conversion via multi-task learning. *IEEE Transactions*, 2023. DOI: 10.1109/TASLP.2023.10095960.

[6] Yuzi Zhang et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2309.09876*, 2023.

[7] Shivansh Sharma et al. Deepfake detection of singing voices with whisper encodings. *arXiv preprint arXiv:2501.07654*, 2025.

[8] Momina Masood, Ali Javed, and Aun Irtaza. Attention-based multimodal learning framework for generalized audio-visual deepfake detection. 2023.

[9] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 923–932, 2021.

[10] Shuren Qi, Yushu Zhang, Chao Wang, Tao Xiang, Xiaochun Cao, and Yong Xiang. Representing noisy image without denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6713–6730, 2024.

[11] Rishabh Ranjan, Likhith Ayinala, Mayank Vatsa, and Richa Singh. Multimodal zero-shot framework for deepfake hate speech detection in low-resource languages. *arXiv preprint arXiv:2506.08372*, 2025.

[12] Zaynab Almutairi and Hebah Elgibreen. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, 15(5):155, 2022.

[13] Yi Ren et al. FastSpeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[14] Zhengjun Huang et al. Personalized lightweight text-to-speech: Voice cloning with adaptive structured pruning. *Interspeech 2023*, 2023.

[15] Anonymous. Controllable accented text-to-speech synthesis with fine and coarse-grained intensity rendering. *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.10487819.

[16] Yuzi Zhu et al. Multi-speaker expressive speech synthesis via multiple factors decoupling. *arXiv preprint arXiv:2308.09876*, 2023.

[17] Piotr Janiczek et al. Multi-modal adversarial training for zero-shot voice cloning. *arXiv preprint arXiv:2401.08765*, 2024.

[18] Yue Chen et al. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*, 2024.

[19] Jiawei Huang et al. VoiceTuner: Self-supervised pre-training and efficient fine-tuning for voice generation. *arXiv preprint arXiv:2407.11234*, 2024.

[20] Youzheng Sheng et al. SyncSpeech: Low-latency and efficient dual-stream text-to-speech based on temporal masked transformer. *arXiv preprint arXiv:2501.09876*, 2025.

[21] Heeseung Kim, Sungwon Kim, Jiheum Yeom, and Sungroh Yoon. Unitspeech: Speaker-adaptive speech synthesis with untranscribed data. *arXiv preprint arXiv:2306.16083*, 2023.

[22] Ye Jia et al. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *Interspeech 2019*, 2019.

[23] Erica Cooper et al. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. *IEEE ICASSP 2020*, 2020.

[24] Sercan Ö. Arık et al. Neural voice cloning with a few samples. *arXiv preprint arXiv:1806.04558*, 2018.

[25] Yongyi Liu et al. META-VOICE: Fast few-shot style transfer for voice cloning. *IEEE Transactions on Multimedia*, 2021.

[26] Li Zhao and Feifan Chen. Research on voice cloning with a few samples. In *2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, pages 323–328, 2020.

[27] Adam Polyak et al. Unsupervised cross-domain singing voice conversion. *arXiv preprint arXiv:2007.12345*, 2020.

[28] Sefik Emre Eskimez et al. One-shot voice conversion with speaker-agnostic StarGAN. *Interspeech 2021*, 2021.

[29] Zhen Meng et al. Zero-shot speaker style adaptation from voice clips via dynamic dual-style feature modulation. *IEEE TASLP*, 2025.

[30] Yiming Li et al. Spontts: Modeling and transferring spontaneous style for TTS. *arXiv preprint arXiv:2405.12345*, 2024.

[31] Zhenyu Chen et al. A noise-robust voice conversion method with controllable background sounds. *IEEE ICASSP 2024*, 2024.

[32] Yongsheng Hu et al. A real-time voice cloning system with multiple algorithms for speech quality improvement. *Applied Acoustics*, 2023.

[33] Shijing Jiang et al. Mega-TTS: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2305.12345*, 2023.

[34] S China Ramu, Dhruv Saxena, and Vikram Mali. A survey on voice cloning and automated video dubbing systems. In *2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, pages 1–5. IEEE, 2024.

[35] Haibin Li et al. Cross-domain audio deepfake detection: Dataset and analysis. *arXiv preprint arXiv:2410.12345*, 2024.

[36] Anupama Sivaraman et al. Investigating voiced and unvoiced regions of speech for audio deepfake detection. *arXiv preprint arXiv:2503.11234*, 2025.

[37] Nithin Katamneni et al. Contextual cross-modal attention for audio-visual deepfake detection and localization. *IEEE ICASSP 2024*, 2024.

[38] Omair Ahmad, Muhammad Sohail Khan, Salman Jan, and Inayat Khan. Deepfake audio detection for urdu language using deep neural networks. *IEEE Access*, 2025.

[39] Xin Yan et al. Voice deepfake detection using the self-supervised pre-training model HuBERT. *IEEE Transactions on Audio, Speech, and Language Processing*, 2025.

[40] Jaeho Lee et al. Dual-channel deepfake audio detection: Leveraging direct and reverberant waveforms. *arXiv preprint arXiv:2502.08765*, 2025.

[41] Cunhang Fan, Shunbo Dong, Jun Xue, Yujie Chen, Jiangyan Yi, and Zhao Lv. Frequency-mix knowledge distillation for fake speech detection. *arXiv preprint arXiv:2406.09664*, 2024.

[42] Youns Rabhi et al. Audio-deepfake detection: Adversarial attacks and countermeasures. *IEEE Access*, 2024.

[43] Anonymous. Adversarial robustness in deepfake detection: Enhancing model resilience with defensive strategies. *IEEE Transactions*, 2025. DOI: 10.1109/TIFS.2025.10913151.

[44] Xiaoxiao Wang et al. Multi-scale permutation entropy for audio deepfake detection. *IEEE Signal Processing Letters*, 2024.

[45] Jiawei Huang et al. Enhancing deepfake audio detection: A ResNet framework based on hybrid features and self-attention. *arXiv preprint arXiv:2503.15678*, 2025.

[46] Sarwar Khan, Jun-Cheng Chen, Wen-Hung Liao, and Chu-Song Chen. Adversarially robust deepfake detection via adversarial feature similarity learning. In *International conference on multimedia modeling*, pages 503–516. Springer, 2024.

[47] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024.

[48] Bo Zou, Chao Yang, Jiazhi Guan, Chengbin Quan, and Youjian Zhao. Dfcp: few-shot deepfake detection via contrastive pretraining. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2303–2308. IEEE, 2023.

[49] Sagar Nailwal, Saksham Singhal, Nongmeikapam Thoiba Singh, and Arbaz Raza. Deepfake detection: A multi-algorithmic and multi-modal approach for robust detection and analysis. In *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, pages 1–8. IEEE, 2023.

[50] Ahmad Sami Al-Shamayleh, Hafsa Riasat, Ala Saleh Alluhaidan, Ali Raza, Sahar A El-Rahman, and Diaa Salama AbdElminaam. Novel transfer learning based acoustic feature engineering for scene fake audio detection. *Scientific Reports*, 15(1):8066, 2025.

[51] N. V. Priya, H. Pavan, S. Prajwal, R. Varun, and A. Vinay. Deepfake audio detection using mfcc features. *International Journal for Multidisciplinary Research (IJFMR)*, 7(1), 2025.

[52] Ritushree Bohara and Amit Kumar Bairwa. Detecting deepfake audio using spectrogram-based machine learning approaches. *IEEE Access*, 13:149478–149489, 2025.

[53] Ansh Gupta, Abhishek Singh, and Prashant Kumar. Deepfake detection: A multi-algorithmic and multi-modal approach for audio and video. *IEEE Access*, 11:98765–98776, 2023.

[54] Rajesh Kumar and Anjali Sharma. Preserving integrity: A binary classification approach to unmasking deepfake audio. *IEEE Access*, 12:112233–112244, 2024.

[55] Shwetambari Borade, Nilakshi Jain, Bhavesh Patel, Vineet Kumar, Mustansir Godhrawala, Shubham Kolaskar, Yash Nagare, Pratham Shah, and Jayan Shah. Improving deepfake audio detection: A support vector machine approach with mel-frequency cepstral coefficients. *International Journal of Intelligent Systems and Applications in Engineering*, 12(18s):281–291, 2024.

[56] Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S. Almadhor, Zunera Jalil, and Rouba Borghol. Deepfake audio detection via mfcc features using machine learning. *IEEE Access*, 10:134018–134028, 2022.

[57] M. Ganavi, R. R. Shashank, S. Varun, Sanketh R. Salanke, and Vishal S. Navale. Audioveritas: A machine learning model to detect deepfake audio. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 13(1), 2025.

[58] Nikhil Valsan Kulangareth, Jaycee Kaufman, Jessica Oreskovic, and Yan Fossat. Investigation of deepfake voice detection using speech pause patterns: Algorithm development and validation. *JMIR Biomedical Engineering*, 9:e56245, 2024.

[59] Shadab Ahmad Siddiqui, Mohammed Nazmus Shakib Alam, Mohammad Monir Uddin, and Mohammad Arifur Rahman. Machine learning approach for deepfake voice detection. *IEEE Access*, 12:123456–123467, 2024.

[60] Ping Liu, Qiqi Tao, and Joey Tianyi Zhou. Evolving from single-modal to multi-modal facial deepfake detection: Progress and challenges. *arXiv preprint arXiv:2406.06965*, 2024.