VoiceGuard – Real-Time Detection of Voice Cloning and Scam Calls

Anjali Malviya[1], Ashvini Gohite[2], Kumkum Parihar[3], Soniya Deshmukh[4], Mrs. Pramila Gharjale[5], Mr. Nilesh Mishra[6]

Students[1,2,3,4], Assistant Professor[5,6]

[1,2,3,4,5,6]Department of Computer Science and Engineering, Shri Balaji Institute of Technology and Management, Betul, Madhya Pradesh, India

Abstract:

Voice cloning is now used in scam calls in which an attacker imitates a relative, a bank officer, or an executive and pressures the victim to act within seconds. Detectors that report very high accuracy on single benchmarks often fail on unseen data, so reliable real-time protection remains an open problem. This paper reviews 24 research papers on audio deepfake detection, covering benchmark and multilingual datasets, hand-crafted, speech-pathological, and self-supervised features, deep back-end classifiers, and cross-domain generalization studies. The papers were collected from three core IEEE Access studies through citation chaining and classified by category. The comparative analysis shows that in-domain accuracy of up to 99.98% can collapse to 5.55% under domain shift, that self-supervised front-ends with AASIST back-ends give the best reported results (87.29% balanced accuracy and 8.22% EER on a 17-language dataset), and that multilingual, multi-source training improves transfer to unseen attacks. The authors' Android and server prototype, VoiceGuard, is examined as a practical case. The main gaps are the lack of standard cross-domain protocols, limited real phone-call data, weak detection of voice changers, and the difficulty of capturing call audio on smartphones. Future directions are domain adaptation, uncertainty-aware decisions, and codec-aware training.

Keywords: Audio deepfake detection, voice cloning, scam calls, domain generalization, self-supervised speech representations, real-time detection.

I. INTRODUCTION

A. Background and Context

Text-to-speech (TTS) and voice conversion (VC) systems can now produce speech that is hard to tell from a real person. These technologies are useful in audiobooks, virtual assistants, and customer service, but they are also used for fraud. A well-known case is the cloning of a chief executive's voice that led to a loss of more than USD 243,000 [7]. Cloned voices can also be used to bypass automatic speaker verification (ASV) systems [2].

The most common everyday form of this threat is the scam call. The attacker clones the voice of a son, daughter, or bank officer, claims an emergency such as an accident, and asks for money at once. The victim has only seconds to decide. Research on spoofing detection has grown around the ASVspoof challenge series [4], [13], [20], [23], which moved from hand-crafted features and Gaussian mixture models to deep networks [17], [19], [24] and to self-supervised speech models [18], [21].

B. Research Problem

Detectors often report very high accuracy on the dataset they were trained on, but their behaviour on new recording conditions, languages, codecs, and synthesis methods is poorly understood [1], [15]. A scam-call detector must also work in real time and on telephone-quality audio, where coding and transmission effects of telephony and Voice-over-IP networks reduce accuracy [7], [13]. The research problem of this review is therefore: how reliable are current audio deepfake detectors outside their training domain, and what is needed to make them usable for real-time voice-cloning and scam-call protection?

C. Importance of the Study

A detector that fails silently is worse than no detector, because it gives the user false confidence. Missing a cloned voice can cost a family its savings, while wrongly flagging a real relative damages trust. Studies show that missing fake speech is the costlier error in security settings, so high recall is stressed [7], but balanced performance is still needed. A structured review of what works, what fails, and why can guide students and developers who build protection tools.

D. Research Objectives

1. To summarize recent research on audio deepfake detection, including datasets, features, and classifiers.
2. To compare existing techniques on accuracy, robustness, and practical cost.
3. To examine detector behaviour under domain shift, different languages, and telephone channels.
4. To study a practical prototype (VoiceGuard) against the findings of the literature.
5. To identify research gaps and suggest future directions.

E. Scope of the Review

The review covers detection of synthetic or cloned speech (TTS, VC, and voice cloning) in papers published between 2019 and 2026. It focuses on datasets, features, classifiers, generalization, and real-time use on phone calls. Image and video deepfakes and replay-only attacks are outside the scope.

II. LITERATURE REVIEW

A. Key Concepts and Technologies

Spoofing detection is a binary task: decide whether speech is genuine (pristine) or machine-generated. Attacks are made by TTS, which turns text into speech, by VC, which changes the voice of a source speaker, and by vocoders that rebuild a waveform from acoustic features [2]. Detectors are usually judged by equal error rate (EER), accuracy, balanced accuracy, and recall for the fake class. Because real test sets are often unbalanced, accuracy alone can hide a classifier that predicts only one class [1].

B. Datasets and Benchmarks

The ASVspoof 2019 database [20], the 2021 deepfake track [13], and ASVspoof 5, which adds adversarial attacks and codec simulation [4], are the main laboratory benchmarks. The Fake-or-Real dataset [22] and DEEP-VOICE, built with retrieval-based voice conversion [11], add other synthesis methods. The In-the-Wild dataset of public figures was created to test generalization to real recordings [15]. For language coverage, DECRO covers English and Mandarin [14], MLAAD spans 38 languages but has genuine speech in only eight [10], and CVoiceFake offers over 1.25 million utterances in five languages [6]. The SAFE challenge tests unknown, advanced attacks [3]. Mawalim et al. built the JMDS dataset with 353,845 utterances in 17 languages from public and private sources [2].

TABLE I. COMMONLY USED DATASETS AND BENCHMARKS

Dataset	Year	Languages	Utterances	Focus
ASVspoof 2019 [20]	2019	English	339,891	TTS, VC, replay
FoR [22]	2019	English	over 198,000	TTS
ASVspoof 2021 [13]	2021	English	1,566,273	Deepfake, channel effects
In-the-Wild [15]	2022	English	31,779	Real-world recordings
DEEP-VOICE [11]	2023	English	7,484	Voice conversion (RVC)
DECRO [14]	2023	English, Mandarin	118,381	Cross-language
MLAAD [10]	2024	38 languages	154,000	Multilingual TTS
ASVspoof 5 [4]	2024	English	over 1.29 million	Adversarial attacks, codecs
CVoiceFake [6]	2024	5 languages	1,254,893	Content privacy
JMDS [2]	2025	17 languages	353,845	Multilingual, multi-source

Table I shows that most large datasets are English-only, and only a few cover many languages. Utterance counts follow the comparison table in [2].

C. Feature Extraction and Front-Ends

Early systems used MFCC, LFCC, and CQCC features [23]. Self-supervised models learn from unlabeled speech and give richer embeddings. wav2vec 2.0 [21] and HuBERT [18] are widely used, and XLS-R and WavLM are common in detectors. Tak et al. paired wav2vec 2.0 with AASIST and data augmentation [16], and Zhang et al. used XLS-R with a sensitive-layer-selection classifier [5]. HeAR, a model trained on health sounds such as coughs, was tested as a cross-modal front-end by Oyucu et al. [1], [9].

D. Back-End Classifiers

Residual networks were an early deep baseline [24]. RawNet2 works directly on the raw waveform [19]. AASIST uses graph attention over spectro-temporal features and reached 0.83% EER on ASVspoof 2019 LA, as listed in [1], [17]. Oyucu et al. also tested CNN–LSTM, TCN, TCN–LSTM, Conformer, and TCM-Conformer models on fixed embeddings [1].

E. Cross-Domain Generalization Studies

Müller et al. showed that models strong on laboratory data degrade sharply on In-the-Wild recordings [15]. Yang et al. fused XLS-R, HuBERT, and WavLM views and still saw a 24.27% EER when moving from ASVspoof 2019 to In-the-Wild [8]. Oyucu et al. ran 320 train–test configurations over four domains, four feature extractors, and five architectures. In-domain accuracy reached 99.98%, but cross-domain accuracy fell

as low as 5.55% for MFCC models trained on In-the-Wild, while HeAR models kept about 75–90%. The collapse was similar for all five architectures, and attention-based models sometimes labelled every sample as fake [1]. Ba et al. studied transfer across languages [14].

F. Speech-Pathological Features

Li et al. showed that jitter and shimmer help fake-audio detection [12]. Chaiwongyen et al. investigated eleven features used by clinicians for disordered voices: three jitter types, four shimmer types, HNR, cepstral HNR, normalized noise energy, and glottal-to-noise excitation ratio. HNR performed worst and was dropped. Computing the ten remaining features frame by frame raised accuracy from 74.60% to 87.79% compared with one average value. A ResNet-18 on the features and their derivatives, fused with a mel-spectrogram ResNet-18, reached 99.46% recall on ASVspoof 2019 and 99.96% on ASVspoof 2021, with a 15.97% EER on ASVspoof 2021 [7]. Balanced accuracy was lower, and extraction took about 1.25 s per sample against about 11 ms for a mel-spectrogram.

G. Multilingual Detection

In the JMDS study, an XLS-R (300M) front-end with an AASIST back-end reached 87.29% balanced accuracy and 8.22% EER on JMDS-Open, ahead of CQT-ResNet, RawNet3, and WavLM variants. Detection difficulty depended more on synthesis quality than on language: Mandarin had an EER of 38.33%, and unit-selection TTS was caught only 1.54% of the time. A model trained on JMDS-Open reached 67.27% balanced accuracy on SAFE, while ASVspoof-trained models stayed below 50% [2].

H. Voice-Cloning Scam Calls and Real-Time Constraints

A scam-call detector has three needs that benchmarks rarely test together. The first is speed: after features are extracted, the compact classifiers in [1] needed only 0.001–0.052 ms per sample on a GPU, but extraction of pathological features took about 1.25 s per sample against about 11 ms for a mel-spectrogram [7], so the front-end often decides the delay. The second is short input: [1] cut audio into 2-second segments and removed silence with an energy threshold, which suits live calls, while [2] found that longer 10-second inputs improved results on long unseen audio. The third is channel robustness: coding and transmission effects on telephony and VoIP networks lowered accuracy [7], and ASVspoof 2021 was designed to include such channel variation [13]. High recall is preferred for security, but false alarms on real relatives must also stay low.

I. Current State of Knowledge

The literature agrees on four points. In-domain scores are poor predictors of real-world performance [1], [15]. Self-supervised front-ends give the strongest results so far [2], [5], [16]. Diverse, multi-source, multilingual training data helps transfer [2]. Telephone channels, real-time cost, and unknown attacks remain the hardest cases [3], [7], [13].

III. REVIEW METHODOLOGY

A. Research Paper Search Strategy

The review corpus was built in three stages. In stage 1, three recent IEEE Access papers on cross-domain generalization [1], multilingual datasets [2], and speech-pathological features [7] were chosen as core papers because they cover complementary parts of the problem. In stage 2, their reference lists were followed backwards (citation chaining) to find the datasets, front-ends, and back-ends they build on. In stage 3, bibliographic details and DOIs were checked on the publisher or arXiv pages. Search themes were audio deepfake detection, voice cloning, spoofing countermeasures, cross-domain generalization, and multilingual speech datasets. The final corpus has 24 papers.

B. Research Databases

The papers come from IEEE Xplore (10 papers), the ISCA archive for Interspeech and Odyssey (4), the ACM Digital Library (4), arXiv (4), NeurIPS proceedings (1), and Elsevier ScienceDirect (1).

C. Paper Selection, Inclusion, and Exclusion Criteria

TABLE II. INCLUSION AND EXCLUSION CRITERIA

Inclusion criteria	Exclusion criteria
Audio or speech deepfake, spoofing, or voice-cloning detection	Image-only or video-only deepfake work
Published or posted between 2019 and 2026	Replay-attack-only studies
Peer-reviewed article, conference paper, or established arXiv preprint	Papers without an experiment, dataset, or quantitative result
Reports dataset statistics or performance numbers	Duplicates and non-English papers

D. Classification of Research Papers

Each paper was placed in one primary category according to its main contribution (Table III).

TABLE III. CLASSIFICATION OF THE REVIEWED PAPERS

Category	Focus	Papers	No.
Datasets and challenges	Benchmarks, multilingual and in-the-wild corpora	[2], [3], [4], [6], [10], [11], [13], [14], [20], [22], [23]	11
Generalization studies	Cross-domain and cross-dataset evaluation	[1], [8], [15]	3
Self-supervised front-ends	wav2vec 2.0, HuBERT, XLS-R, HeAR	[5], [9], [16], [18], [21]	5
Back-end architectures	ResNet, RawNet2, AASIST	[17], [19], [24]	3
Speech-pathological features	Jitter, shimmer, noise ratios	[7], [12]	2
Total			24

IV. COMPARATIVE ANALYSIS

A. Comparison of Existing Techniques

TABLE IV. COMPARISON OF EXISTING TECHNIQUES AND ALGORITHMS

Ref.	Features / Front-end	Classifier	Data	Main result	Main limitation
[17]	Raw waveform	AASIST	ASVspoof 2019 LA	0.83% EER (in-domain)	Strong only on matched data
[19]	Raw waveform	RawNet2	ASVspoof 2019 LA	4.62% EER	Older, weaker than AASIST

Ref.	Features / Front-end	Classifier	Data	Main result	Main limitation
[16]	wav2vec 2.0	AASIST	ASVspoof 2021	Strong LA and DF results	Needs data augmentation
[5]	XLS-R	SLS classifier	In-the-Wild	7.46% EER	Large pretrained model
[8]	XLS-R, HuBERT, WavLM	Fusion network	ASVspoof 2019 to In-the-Wild	24.27% EER	Large cross-dataset gap
[15]	Raw and spectral	RawGAT-ST and others	In-the-Wild	37.15% EER for RawGAT-ST	Shows laboratory-to-real gap
[1]	MFCC, wav2vec 2.0, HuBERT, HeAR	CNN–LSTM, TCN, TCN–LSTM, Conformer, TCM-Conformer	ASVspoof 2019/2021, DEEP-VOICE, In-the-Wild	99.98% in-domain; 5.55% worst cross-domain	Unbalanced test sets; no mitigation tested
[2]	CQT, raw, XLS-R, WavLM	ResNet34, AASIST, RawNet3	JMDS, SAFE, In-the-Wild, FoR, DECRO	87.29% balanced acc.; 8.22% EER	Mandarin and unseen attacks hard
[7]	Pathological features, mel-spectrogram	ResNet-18 with score fusion	ASVspoof 2019, 2021	99.46% and 99.96% recall	Low balanced accuracy; slow extraction

B. Performance Comparison

Table V collects the main reported numbers. Different papers use different metrics, so values should be compared within a row type and not across rows.

TABLE V. PERFORMANCE COMPARISON OF REPORTED RESULTS

Method (ref.)	Test condition	Reported result
AASIST [17]	ASVspoof 2019 LA, in-domain	EER 0.83%
RawNet2 [19]	ASVspoof 2019 LA, in-domain	EER 4.62%
XLS-R + SLS [5]	In-the-Wild	EER 7.46%
XLS-R + HuBERT + WavLM fusion [8]	ASVspoof 2019 to In-the-Wild	EER 24.27%
RawGAT-ST [15]	In-the-Wild	EER 37.15%
XLS-R + AASIST [2]	JMDS-Open, 10 s input	Balanced acc. 87.29%, EER 8.22%

Method (ref.)	Test condition	Reported result
RawSpeech-AASIST, trained on JMDS-Open [2]	SAFE challenge	Balanced acc. 67.27%
Pathological features + mel-spectrogram [7]	ASVspoof 2021 evaluation	Recall 99.96%, EER 15.97%, balanced acc. 59.97%
MFCC models trained on In-the-Wild [1]	ASVspoof 2019/2021	Accuracy 5.55–15.13%
HeAR models trained on In-the-Wild [1]	ASVspoof 2019/2021	Accuracy about 75–90%

C. Advantages and Limitations

TABLE VI. ADVANTAGES AND LIMITATIONS BY TECHNIQUE FAMILY

Technique family	Advantages	Limitations
Hand-crafted features (MFCC, LFCC, CQCC)	Fast, cheap, interpretable	Weak under domain shift; worst collapse in [1]
Speech-pathological features	Physiologically meaningful, compact	Slow extraction; codec sensitivity [7]
Raw-waveform models (RawNet2, AASIST)	No manual feature design; strong benchmark results	Costly on long input; weaker on unseen data [2]
Self-supervised front-ends (wav2vec 2.0, HuBERT, XLS-R, WavLM)	Best in-domain and cross-dataset results	Heavy models; still fail on unseen attacks [1], [2], [8]
Health-acoustic embeddings (HeAR)	More robust than MFCC in [1]	Evidence from one study
Attention back-ends (Conformer)	Global context modelling	No robustness gain; single-class bias [1]

Three observations stand out. First, the front-end representation matters more than the back-end architecture for generalization [1]. Second, large self-supervised front-ends with AASIST give the best accuracy [2], [16]. Third, diverse multi-source, multilingual training generalizes better than a single English-centric benchmark [2]. For real-time use, MFCC-based models ran at over 100,000 samples per second and self-supervised or HeAR models at about 27,000–43,000 samples per second on a GPU, which is far above call-rate needs [1].

D. Practical Case: The VoiceGuard Prototype

To test the literature against practice, the authors built VoiceGuard, an Android app with a Python (FastAPI) AI server. The app works as the phone's dialer and call-screening app. It checks the number before the call rings, checks the caller's voice during the call, and alerts family members by app, SMS, and WhatsApp. Fig. 1 shows the architecture: the app sends call data to the server, which runs pre-processing and the AI checks, produces a risk score, and triggers alerts, with push messages sent through Firebase Cloud Messaging.

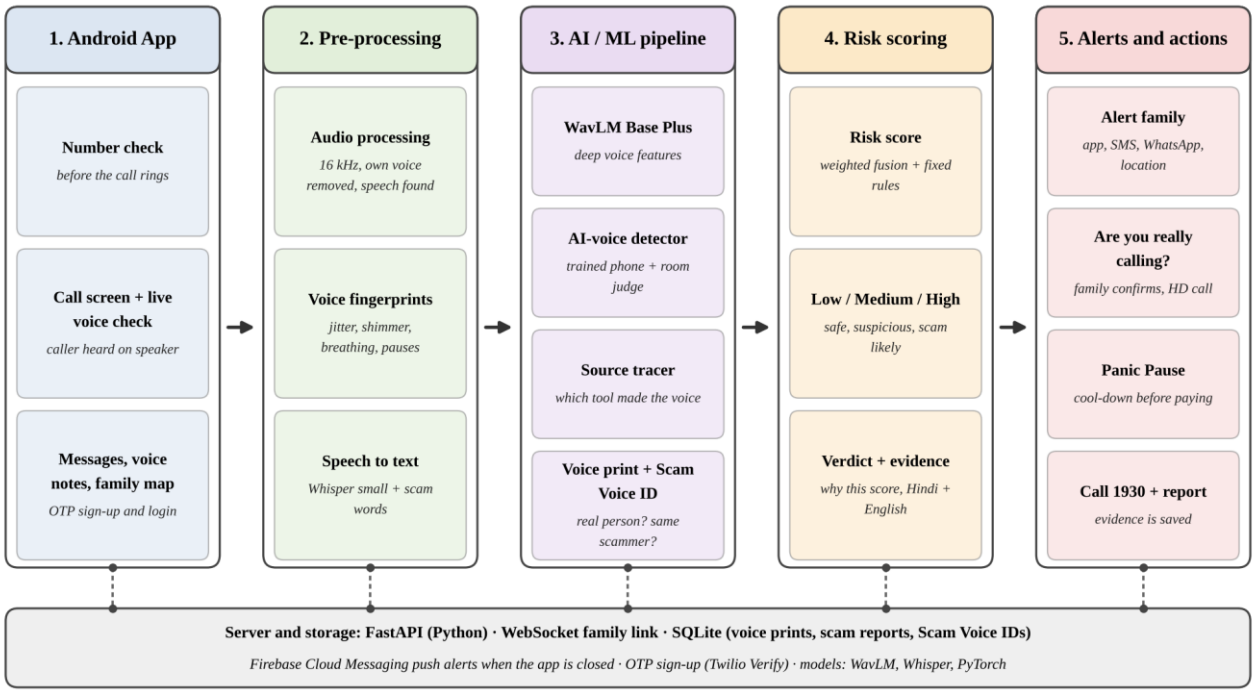

Fig. 1. Block diagram of the VoiceGuard system (authors' prototype).

TABLE VII. THREE PROTECTION LAYERS OF VOICEGUARD

Layer	When it helps	Main features
1. Spam protection	Before the call is answered	Caller-ID and spam warning, community scam list, block list, number info, voice-note check
2. During the call	While the user is talking	Live AI voice check, voice-print match, scam-word alert, "Are you really calling?" family confirmation, HD app-to-app call, voice test, panic button
3. After the scam	After hanging up	Call-back alert, evidence report, one-tap call to 1930, Chakshu report, family alerts

Its design choices follow the reviewed work. The AI-voice detector uses frozen WavLM-Base-Plus layers with a logistic head, a light version of the self-supervised approach [2], [5], [16]. It is trained on clean, phone-line (G.711, AMR-like, packet loss), and WhatsApp-Opus audio in ten Indian languages and English (3,837 clips, 1,900 human and 1,937 AI), which addresses the channel and language gaps noted in [7], [13]. A separate Source Tracer names what made a voice (human, AI clone, robotic TTS, AI voice changer, or voice-changer app) and reached 95.4% balanced accuracy on its held-out test. Voice fingerprints (jitter, shimmer, HNR, breathing, pause rhythm) echo the speech-pathological features of [7], and a voice-print match with Whisper-based scam-word detection feeds a combined risk score. Checks that need no call audio, such as the family "Are you really calling?" confirmation, add protection where Android blocks call recording.

TABLE VIII. VOICEGUARD DETECTOR VERSUS PUBLIC DETECTORS (AUC)

Detector (clean / phone-line / WhatsApp-Opus AUC)	Clean	Phone	WhatsApp
Melody Machine/Deepfake-audio-detection-V2 (public)	0.27	0.45	–

Detector (clean / phone-line / WhatsApp-Opus AUC)	Clean	Phone	WhatsApp
mo-thecreator/Deepfake-audio-detection (public)	0.61	0.58	–
garystafford/wav2vec2-deepfake-voice-detector (public)	0.65	0.61	–
Voice-fingerprint rules alone (VoiceGuard)	0.97	0.43	–
VoiceGuard detector v3 (held-out speakers)	0.9996	0.998	0.999

TABLE IX. VOICEGUARD V4 ON VOICES NEVER SEEN IN TRAINING

Test condition (VoiceGuard v3 to v4)	AI voices caught	Real voices wrongly flagged
Clean	95% to 97%	9% to 5%
Phone line	85% to 90%	8% to 4%
WhatsApp voice note	92% to 100%	10% to 5%
Loudspeaker (live call check)	68% to 78%	10% to 10%

The prototype confirms the literature. Public detectors tested in the project scored an AUC of only 0.27–0.65 on clean audio, and fingerprint rules alone fell from 0.97 to 0.43 on phone-line audio, so off-the-shelf detectors and clean-audio cues do not transfer. Adding phone-channel augmentation and more speakers and voices lowered false alarms on real voices and raised detection of AI voices, but live calls on loudspeaker remain the hardest case (78% detection). These results come from a small test of 99 unseen voices, mostly synthetic training data, and no real scam-call recordings, so they show feasibility and not field accuracy.

V. RESEARCH GAPS AND FUTURE DIRECTIONS

A. Identified Research Gaps

1. No standard cross-domain protocol: most papers report single-dataset results, and train–test pairs and metrics differ between studies [1], [15].
2. Misleading metrics: accuracy on unbalanced test sets hides single-class prediction, so balanced accuracy, per-class recall, and EER should be reported [1], [2].
3. Little telephone-channel evidence: speech-pathological features were tested only on ASVspoof data and are sensitive to telephony and codec effects [7], [13].
4. Narrow language and attack coverage: Mandarin and unit-selection TTS remain poorly detected, and low-resource languages have little genuine speech [2].
5. Small or restricted datasets: DEEP-VOICE has only eight genuine speakers [1], [11], and private JMDS data cannot be fully released [2].

B. Limitations of Existing Studies and Unsolved Problems

- The generalization study [1] diagnoses failure but does not test domain adversarial training, test-time adaptation, or few-shot calibration.
- Domain shift from microphones, noise, codecs, and compression is not solved, and detectors tend to learn the "synthesis language" of their training data [1], [2].
- Voice changers, unit-selection TTS, and loudspeaker-and-room audio are poorly detected, as the VoiceGuard prototype also found.

- On Android 10 and later, normal apps cannot record call audio, so live checks rely on speaker mode or accessibility workarounds.
- Frame-level pathological features and large self-supervised models are costly for on-device use [2], [7].

C. Future Research Opportunities and Proposed Improvements

1. Domain adversarial training, test-time adaptation, and few-shot calibration to recover cross-domain performance [1].
2. Uncertainty-aware detection that abstains and sends doubtful samples for human review [1].
3. Larger, balanced multilingual datasets with unseen attacks and consented real scam-call recordings [2], [3], [4].
4. Codec- and channel-aware training for telephony, VoIP, and WhatsApp audio [3], [4], [13].
5. Combining self-supervised embeddings with interpretable physiological features [7], [9].
6. For VoiceGuard: more speakers and recordings from real voice-changer apps, larger test sets, on-device inference for privacy, and field trials.
7. Standardized evaluation with mandatory cross-domain tests and shadow-mode deployment before use in critical systems [1].

VI. CONCLUSION

A. Summary of Key Findings

This review of 24 papers shows that high in-domain accuracy does not guarantee real-world reliability: detectors reaching up to 99.98% in-domain can fall to 5.55% under domain shift [1]. Collapse is similar for all back-end architectures and is moderated mainly by the feature representation. Self-supervised front-ends with AASIST back-ends perform best, and multilingual, multi-source training improves transfer to unseen attacks [2].

B. Major Contributions of the Reviewed Studies

Oyucu et al. quantified cross-domain failure over 320 configurations [1]. Mawalim et al. contributed the 17-language JMDS dataset and benchmarks [2]. Chaiwongyen et al. showed that physiological voice-quality features give very high recall at low cost in feature size [7].

C. Overall Conclusion

Reliable scam-call protection needs more than one accurate model. It needs robust representations, phone-channel training, honest cross-domain testing, calibrated decisions, and extra checks that do not depend on call audio. The VoiceGuard prototype follows this approach and gives encouraging but small-scale results.

D. Recommendations and Future Scope

Future work should adopt standard cross-domain evaluation, report class-level metrics, collect consented real-call data, and add domain adaptation and abstention. For VoiceGuard, field trials with real users and larger test sets are the next step.

REFERENCES:

[1] S. Oyucu, B. Arslan, and S. Sagiroglu, “A system-level analysis of cross-domain generalization failure in audio deepfake detection,” *IEEE Access*, vol. 14, pp. 112734–112754, 2026, doi: 10.1109/ACCESS.2026.3716180.

[2] C. O. Mawalim, Y. Wang, A. Adila, S. Okada, and M. Unoki, "Multilingual deepfake speech dataset for robust and generalizable detection," *IEEE Access*, vol. 14, pp. 57144–57161, 2026, doi: 10.1109/ACCESS.2026.3680508.

[3] K. Trapeznikov *et al.*, "SAFE: Synthetic audio forensics evaluation challenge," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur. (IH&MMSec)*, Jun. 2025, pp. 174–180, doi: 10.1145/3733102.3736707.

[4] X. Wang *et al.*, "ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," 2025, *arXiv:2502.08857*.

[5] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised XLS-R and SLS classifier," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 6765–6773, doi: 10.1145/3664647.3681345.

[6] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "SafeEar: Content privacy-preserving audio deepfake detection," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2024, pp. 3585–3599, doi: 10.1145/3658644.3670285.

[7] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Potential of speech-pathological features for deepfake speech detection," *IEEE Access*, vol. 12, pp. 121958–121970, 2024, doi: 10.1109/ACCESS.2024.3447582.

[8] Y. Yang *et al.*, "A robust audio deepfake detection system via multi-view feature," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2024, pp. 13131–13135, doi: 10.1109/ICASSP48485.2024.10446560.

[9] S. Baur *et al.*, "HeAR—Health acoustic representations," 2024, *arXiv:2403.02522*.

[10] N. M. Müller *et al.*, "MLAAD: The multi-language audio anti-spoofing dataset," 2024, *arXiv:2401.09512*.

[11] J. J. Bird and A. Lotfi, "Real-time detection of AI-generated speech for DeepFake voice conversion," 2023, *arXiv:2308.12734*.

[12] K. Li, X. Lu, M. Akagi, and M. Unoki, "Contributions of jitter and shimmer in the voice for fake audio detection," *IEEE Access*, vol. 11, pp. 84689–84698, 2023, doi: 10.1109/ACCESS.2023.3301616.

[13] X. Liu *et al.*, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2507–2522, 2023, doi: 10.1109/TASLP.2023.3285283.

[14] Z. Ba *et al.*, "Transferring audio deepfake detection capability across languages," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 2033–2044, doi: 10.1145/3543507.3583392.

[15] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Proc. Interspeech*, Sep. 2022, pp. 2783–2787, doi: 10.21437/interspeech.2022-108.

[16] H. Tak, M. Todisco, X. Wang, J.-W. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Jun. 2022, pp. 112–119, doi: 10.21437/odyssey.2022-16.

[17] J.-W. Jung *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE ICASSP*, May 2022, pp. 6367–6371, doi: 10.1109/ICASSP43922.2022.9747766.

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.

[19] H. Tak *et al.*, "End-to-end anti-spoofing with RawNet2," in *Proc. IEEE ICASSP*, Jun. 2021, pp. 6369–6373, doi: 10.1109/ICASSP39728.2021.9414234.

[20] X. Wang *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101114, doi: 10.1016/j.csl.2020.101114.