Audio Deepfake Detection and Defense: A Systematic Review with Real-Time Framework for Indian and Global Contexts

Sandeep Verma[a,*], Seema Bawa[a], Sachin Kansal[a], Arpit Jain[a], Japneet Singh[a], Kaustubh Singh[a], Diwakar Narayan Sood[a], Shivane Kapoor[a]

[a]*Department of Computer Science and Engineering, Thapar Institute of Engineering & Technology, Patiala-147004, India*

Abstract

Audio deepfakes pose a major security concern, especially in multilingual countries like India, where voice-based fraud surpasses literacy barriers. This work presents a systematic review of methods for detecting AI-generated voice content in Indian languages. Starting from results by Google Scholar, arXiv, and IEEE Xplore from 2020 to 2025, selected studies are based on relevance, performance metrics, and language support. The papers reviewed range from spectrogram-based CNNs to transformer-based architectures, and most works focus on Hindi and English, with very few exploring regional languages. While detection accuracies are often greater than 90%, the majority of models lack generalisation across diverse dialects and real-world conditions. This review highlights existing gaps, including the need for broader language coverage, improved robustness, and practical deployment strategies, while consolidating references for future work related to Indian language audio deepfake detection.

Keywords: Deepfake Detection, Hindi, English, Hinglish, Multimodal AI, Audio Manipulation, Regional Language AI, Systematic Review, PRISMA, Review Article

1. Introduction

These days, in the digital domain, audio, video, and image content is created and shared faster than ever; hence, this multimedia is an effective means of communication and influence
citeyi2023survey
citesarala2024social. However, this development comes with a critical threat: deepfakes synthetically generated media capable of convincingly replicating real human appearance and voices
citeshaaban2025audio. Of these, the most disquieting are the audio deepfakes. Advanced AI models clone voices to synthesize speech that is almost indistinguishable from utterances made by humans
citealmutairi2022review. These counterfeit audio signals not only deceive listeners but also give rise to misinformation
citemittal2024pitch, break public trust, and present threats to national and personal security
citezhang2025audio. Deepfakes can generally be classified as image deepfakes, which are tampered facial imagery, video deepfakes, which are face-swapped or reenacted visuals, and audio deepfakes in voice cloning or speech synthesis. Of these, the audio-based manipulations have become increasingly undetectable, especially in regions with high linguistic diversity, such as India, allowing political manipulation, impersonation fraud, and identity theft. Facial and vocal manipulation has evolved from rule-based models to modern deep learning paradigms. An important development was the introduction of systems such as Video Rewrite and Face2Face - early solutions for visual lip-syncing and reenactment[1]. Further advances in GANs[2], autoencoders, and transformers have enabled the synthesis of highly realistic content on both audio and

*Corresponding author.
Email address: sandeep.verma@thapar.edu (Sandeep Verma)

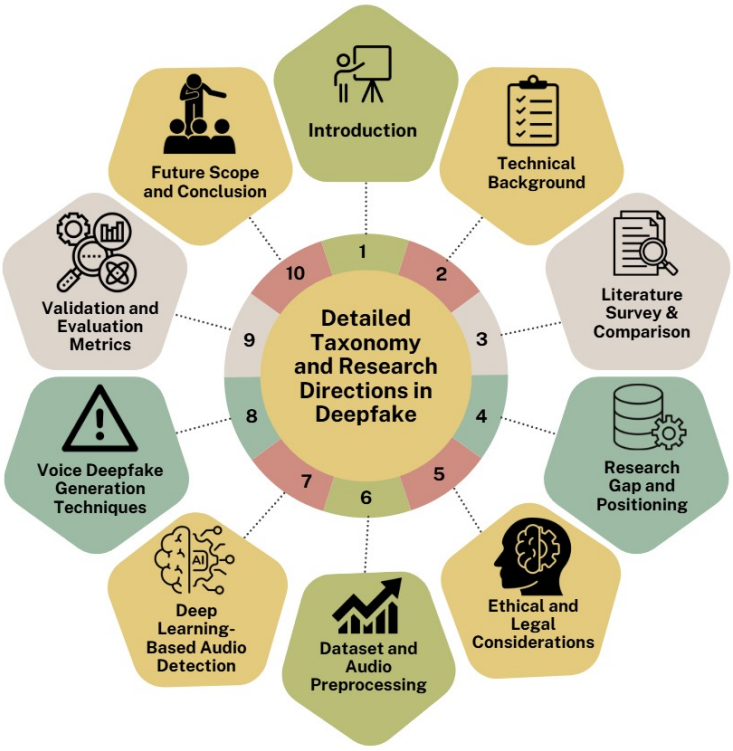

Figure 1: Taxonomy of Deepfake Research Directions. The PRISMA-driven framework covers 10 core areas relevant to multilingual audio deepfake detection.

video domains [3]. With the latest diffusion models and large-scale training pipelines[4], modern deepfake generators can reproduce vocal identity, prosody[5], emotion, and accent with unprecedented accuracy [6].• Recent studies are also emphasizing the development of energy-efficient detection systems [7] that would be scalable for real-time applications. For reliable detection, manual or visual inspection is far from sufficient. Thus, automated detection systems have been developed that range from classical machine learning to advanced deep learning techniques [8] [9]. These have included convolutional neural networks (CNNs) [10], temporal models such as RNNs and GRUs, spectrogram-based classifiers such as ResNet and ResNeXt, and transformer-based architectures for frame-level prediction [11][12][13]. However, practical deployment still faces the challenge of generalization in datasets [14], robustness against adversarial attacks [15], and low-latency real-time inference [16]. This paper has presented a systematic review, following the PRISMA 2020 guidelines, of existing approaches to detect AI-generated audio deepfakes with a special emphasis on the Indian linguistic context.It reviews existing detection methods, points out important gaps, especially on the level of regional languages, and proposes a new multilingual detection system with an optimization for real-time scenarios of fraud cases[17]. Such challenges are addressed in the present work, which is related to real-time voice fraud detection in multilingual environments of India. Unlike existing tools that target either English or Hindi, this approach was customised to detect regional voice-cloning attacks in languages such as Tamil, Marathi, Bengali, and Telugu [18].By leveraging spectral preprocessing, deep learning-based inference, and multilingual on-device optimization, this work offers proactive protection against audio-based impersonation in real-world calls [19]. Figure 1 presents the taxonomy of the reviewed literature.

1.1. Major Contributions

1.2. Major Contributions

- This work presents a comprehensive and updated systematic review of methods for the detection of audio deepfakes [20]. The review covers both classical machine learning and deep learning approaches

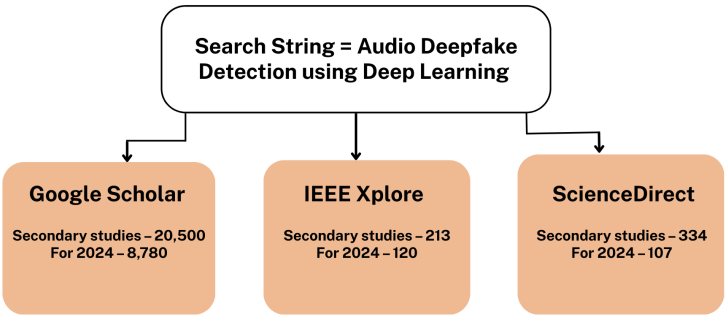

Figure 2: The sources of the previous research articles used in the research.

[21]. Relevant surveys are also included for completeness [4].

- This review follows the PRISMA methodology [20], which in detail describes the inclusion criteria, search strategy, study synthesis, and a PRISMA flow diagram.
- Recent surveys are compared with a few previous works [22]. For example, the study by [22] looks at things in a wider, more multilingual way and covers many technical angles. Another survey, [23], focuses more on under-resourced Indian regional languages. We also looked at [24] as part of this comparison.
- Key benchmark datasets have been looked into [25]. Different detection models like CNNs, RNNs, and transformers are checked for how well they generalise across domains [18]. Some works also try multimodal setups to get better real-time performance and handle adversarial cases more strongly [16].
- Ethical, societal, and legal implications in the Indian context are discussed in [26]. A few works also talk about future ideas for building scalable and privacy-friendly detection systems [27], with a special focus on protecting people who might get affected the most by deepfake misuse [14].

Taxonomy of the Proposed Work

The rest of the paper is set up like this. First, Figure 2 gives a quick overview of the initial literature search we did on "Audio Deepfake Detection using Deep Learning," mainly showing how the research interest has been growing across different academic sources. Then, Section 2 goes over some basic ideas and important concepts related to audio and video deepfakes. After that, Section 3 compares earlier survey papers. Section 4 looks at different detection pipelines and explains how our work fits into what has already been done. Sections 5 and 6 talk about how deepfakes are made and how deep learning methods are used to detect them. Section 8 discusses the datasets and benchmarks that are commonly used in audio-visual deepfake detection. Finally, Section 7 covers the ethical, legal, and social concerns around deepfakes. Section 9 discusses the key research questions, and Section 10 addresses up-and-coming directions and remaining challenges. Finally, Section 11 wraps up the paper with concluding insights and recommendations.

2. Background Overview, Research Questions, and Workflow of Deepfake Audio Detection

Deepfakes can be broadly classified into three key modalities audio, video, and image each posing unique challenges in detection and defense, especially in multilingual, culturally diverse regions such as India as also shown in Figure **??**.

2.1. Audio-deepfakes

Audio-deepfakes involve mimicking or synthesizing human speech using artificial intelligence, posing significant risks such as impersonation frauds and phone scams [28]. Figure 3 presents the PRISMA-based selection process, showing how 385 records were screened and refined to 36 final studies included in the review. This work reviews methods that aim to address these threats, particularly in multilingual contexts.

- Voice cloning uses short samples of a real voice to replicate an individual's speech patterns, enabling attackers to impersonate family members, colleagues, or executives to extract money or sensitive information[29][30].
- Text-to-speech (TTS) converts written text into synthetic speech, which, while beneficial for accessibility and automation, can be misused to create fraudulent calls, alerts, or misinformation[28][31].
- Emotionally conditioned synthetic speech embeds voices with emotional cues, such as urgency or anger, to make fraudulent communications more persuasive[32][33].

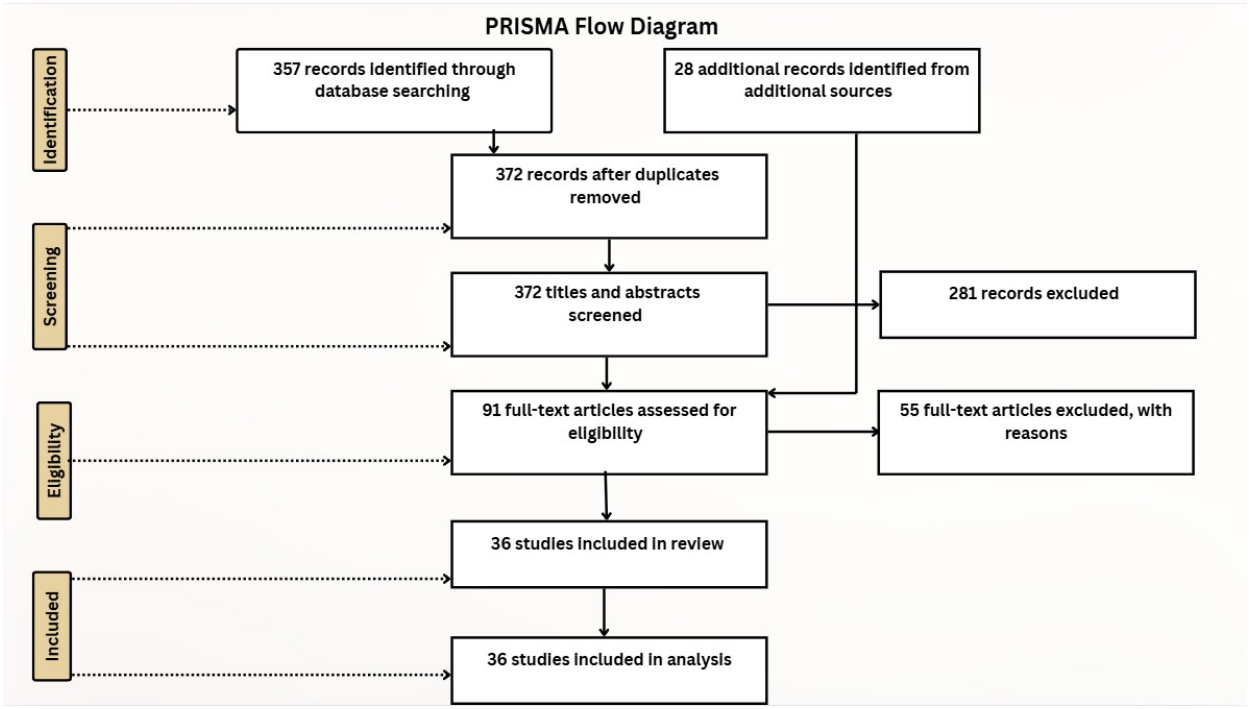

Figure 3: PRISMA Flow Diagram displaying study selection and inclusion process.

2.2. Video-deepfakes

In the case of video deepfakes, facial expressions, movements, or sometimes the identity itself is tampered with using AI. Misinformation, blackmail, and impersonation attacks [34] based on these manipulations are becoming increasingly prevalent. The three primary types of deepfakes—video, image, and audio—are presented in Figure 4. Each of these reflects a different kind of AI media manipulation [35].

- This deals with the replacement of one person's face for another, a technique familiar in fake celebrity videos or cases of identity theft[36].
- Lip-syncing involves the editing of lip movements in accordance with altered audio to create an effect that suggests a person spoke words which were never said [37].

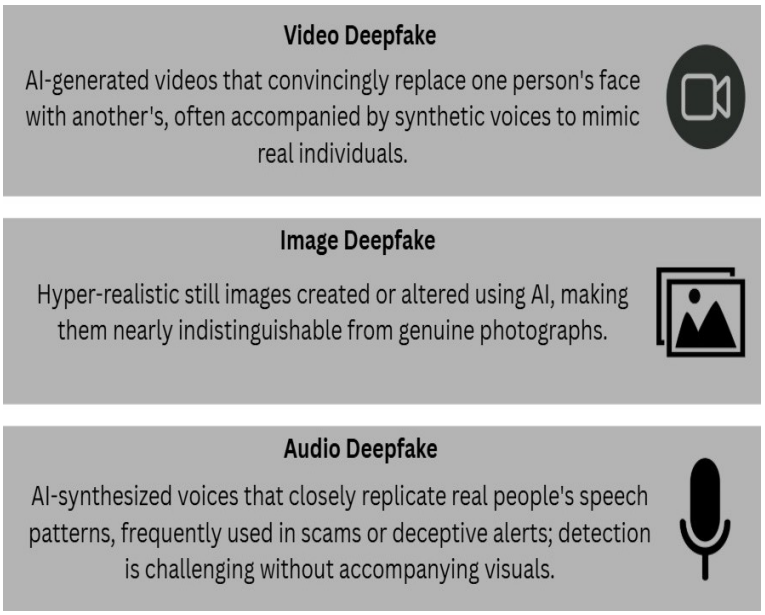

Figure 4: Illustration of deepfake categories, including video, image, and audio variants.

- Facial reenactment enables real-time control of someone's facial appearance through the movements of another actor. This is now commonly used in fabricated interviews or politically motivated media content [38][39].

  avatars are completely artificial humans, created by machine learning models, that are often used to circulate disinformation, conduct scams, or dupe audiences with sham marketing [34].

2.3. Image-deepfakes

Image deepfakes, though static, are very influential and may be damaging. They are often used to fabricate online profiles or alter existing photos in such ways that mislead, deceive, or tarnish reputations [18].

- Synthetic faces are computer-generated portraits of people that do not really exist. Such images can be found in fake social media accounts, false identities, or forged documents [36].
- Face morphing combines two real faces into a new composite identity, which may be exploited to bypass biometric systems, such as face recognition or passport verification checks [27].
- AI-based image manipulation utilises artificial intelligence to alter existing photographs, changing backgrounds, removing people, or inserting fabricated elements in highly realistic and convincing ways [40].

2.4. Research Questions

The review is intended to fill in the gaps on critical aspects of research on audio and video deepfake detection within the Indian context, following the research design according to the PRISMA methodology[5]. The study was informed by the following research questions, formulated prior to the review process:

- **RQ1:** Do current benchmark models generalize effectively to Indian deepfake datasets[11]?
- **RQ2:** Which audio features actually works best for spotting deepfake clues across different Indian languages[32]?

- **RQ3:** How can we make these detection systems run better on mobile devices or other setups where resources are limited[41]?
- **RQ4:** How good are the existing models at spotting deepfakes when the speech is code-mixed[16]?

These research questions shaped the inclusion and exclusion criteria for the study and guided the synthesis of the reviewed literature. The findings and responses to these questions are consolidated and discussed in the concluding sections of this paper.

2.5. Workflow of Deepfake Detection

The detection of audio-deepfakes particularly those generated through text-to-speech (TTS), voice conversion (VC), or voice cloning requires a structured, multi-stage pipeline that combines signal processing, discriminative feature extraction, and deep learning based classification. The following stages outline a generalized detection framework synthesized from the reviewed literature. This description is based on prior research and does not represent an implemented system.

1. Input acquisition: Acquire audio from real-time phone calls, voice messages, or stored recordings in Indian languages, including code-mixed speech.
2. Audio pre-processing:
    - Noise suppression (e.g., spectral gating)
    - Voice activity detection to remove silence and non-speech segments
    - Resampling and amplitude normalization
3. Feature extraction: Extract features sensitive to synthetic-speech artifacts, such as speech pause patterns [42]:
    - Spectral: MFCC, CQCC, LFCC
    - Temporal: Zero-crossing rate, short-time energy
    - Prosodic: Pitch, intonation, speaking rate
    - Time–frequency: Mel-spectrograms, log-power spectrograms
4. Audio-deepfake detection model: Train or fine-tune models using the extracted features:
    - CNN architectures such as ResNet and ResNeXt for spectral-input analysis
    - Transformer-based models such as Wav2Vec 2.0 for fine-grained speech embeddings
    - Entropy-based approaches such as f-InfoED (frame-level latent information entropy)
5. Classification layer: Classify the audio as bona fide or spoofed using:
    - Softmax or sigmoid binary classifiers
    - Siamese networks or contrastive loss such as StacLoss for pairwise learning
6. Post-processing and alerting:
    - Apply decision thresholds to trigger real-time spoofing alerts

This architecture, adapted from state-of-the-art detection pipelines cited in related literature, is intended for robust detection of audio-deepfakes in multilingual and code-mixed speech environments. The description is part of a literature review and is not an implementation..

3. Comparison with Existing Literature Survey

Recent advances in audio- and multimodal-deepfake detection have demonstrated strong performance in controlled environments. However, robustness in real-world conditions remains a significant challenge. This section synthesizes findings from prior studies selected through a structured PRISMA compliant review process. Only peer-reviewed works or preprints from 2020–2025 were considered, with inclusion criteria focusing on dataset usage, linguistic diversity, real-world applicability, and availability of performance metrics.

[1] examines detection strategies based on spectrogram-driven deep learning. [43] explores multimodal frameworks combining audio and video. Despite strong results in controlled settings, both approaches underperform on unseen accents, code-switched speech, and noisy or compressed mobile recordings. [20] attains high accuracy on clean datasets but reports Major decline under varied acoustic profiles. [5] shows the generalization gap in current models. [18] introduces the HAV-DF dataset to incorporate Hindi-language deepfakes, thereby enhancing linguistic diversity in evaluation pipelines.
[11] proposes a hybrid model fusing waveform and spectrogram features. [8] adopts ResNeXt-based backbone for better representation learning. Both are still significantly vulnerable against adversarial audio. [22] shows that feature extraction specifically designed for Indian languages leads to increased performance. This is supported by [44], which shows similar improvements with optimized spectrogram-based features. [15] proposes the JMAD dataset for over-coming limitations in cross-language training. [24] introduces the MLAAD dataset with extended multilingual coverage. [16] contributes MLADDC for diversified audio-deepfake detection scenarios. [45] adds multilingual and multimodal testing support. [41] presents PITCH, which is a real-time tagging protocol using challenge response verification, aligning with real-world operational needs.

Several works address vulnerabilities in compressed social-media voice recordings and code-switched audio, for example, Hinglish, where many models exhibit steep performance drops. [5] informs entropy-based modeling approaches. [46] supports prosody-based analysis modules. Emotional fingerprinting for speaker verification is discussed in [6]. [3] contributes adversarial defense strategies. [14] adds complementary robustness measures. The works by Singh et al. and Kaur et al. have informed the cross-lingual evaluation strategies.

[2] guides low-latency inference design for practical deployment. This direction moves beyond the English-centric bias of most existing datasets, making the evaluation framework more inclusive and realistic.

This work investigates sophisticated multimodal fusion approaches by [11]. [8] explores similar integrations, though both remain susceptible to highly engineered audio-deepfakes
exploiting model blind spots. Following PRISMA guidelines, Table 1 offers a tabulated synthesis of selected studies, summarising their core characteristics, methodologies, and alignment with real-world multilingual audio-deepfake detection objectives.

4. Critical Literature Assessment and Research Positioning

This survey targets the detection of audio deepfakes in Indian languages [21]. Whereas earlier works have tested model performance primarily on clean and controlled datasets [20], several have not considered any real-world issues, including background noise, multilingual code-switching-for example, Hindi–English-and regional accents-which normally happen in Indian phone conversations. The methods reviewed within this work include some designed to combat these practical issues specifically [22]. Table 2 summarizes the earlier work in terms of deep learning architectures, feature types, multilingual capabilities, cross-language generalization, and deployment suitability, with a highlight of ILADEF's strengths in real-time, mobile, and edge/on-device applications.

A significant differentiating factor in the works reviewed is the inclusion of multilingual training and evaluation settings [15]. The work [24] evaluates the generalization performance across languages with several Indian languages, like Hindi, Tamil, Bengali, and Marathi. This can be used to study whether a system trained in one language works effectively for any other language to detect audio deepfakes [16]. Furthermore, there have also been recent studies involving advanced voice cloning systems of high fidelity that generate synthetic speech which is almost indistinguishable from natural, genuine human voices [41].

Table 1: Comparative Summary of Indian Language Audio Deepfake Defense (ILADEF) and Related Works

Work	**Datasets Used**	**Strengths**	**Weaknesses**	**Generalizability**	**Research Gaps Addressed**
Shaaban & Yildirim (2025)[3]	TTS synthetic English/Arabic	CNN spectral DL methods benchmarked	No cross-language experiments	Moderate	Highlights need for dataset diversity
Jin et al. (2025)[11]	Cross-modal spectrogram fusion	Wave-spectrogram aggregation, temporal+frequency integration	GPU heavy, not real-time suitable	Low	Encourages fusion of complementary features
Zhao et al. (2025)[5]	Frame-level entropy features	Generalized latent entropy analysis	Assumes clean inputs; lacks noisy call simulations	Moderate	Frame entropy as new discriminative feature
Tahaoglu et al. (2025)[8]	Spectral ResNeXt, English data	Lightweight spectral ResNeXt architecture	No dialect or multilingual training	Moderate	Efficient CNN with limited language diversity
Ranjan et al. (2025)[2]	Hindi	Detecting hate speech in synthetic audio	Focus on hate, not generic fraud	Medium	Adds content semantics to deepfake studies
Mawalim et al. (2025)[15]	JMAD: 38 language audio deepfakes	Diverse multi-accent corpus	No Hindi specific subtest	Very high	Foundation for multi-accent detection
Kaur et al. (2024)[18]	HAV-DF (Hindi AV deepfake)	First large Hindi audio-video deepfake dataset	Single-language, mono-modal tests	Low	Opens benchmark for Indian regional studies
Purohit et al. (2024)[47]	GGMDDC: Hindi + global GAN/TTS deepfakes	Multilingual balanced dataset	Still early stage corpus	High	Builds broader datasets for multilingual spoof detection
Phukan et al. (2024)[40]	Multilingual Wav2Vec2/PTM speech	PTMs for multi-lang deepfakes	Mostly feature benchmarking	High	Tests low-resource cross-language generalization
Almutairi & Elgibreen (2022)[43]	English/Arabic corpora	Systematic survey of audio deepfake detection algorithms	No multilingual datasets; no experimental models	Moderate: general architectures discussed	Identifies gaps in multilingual research
Ambili & Roy (2022)[23]	Hindi, Tamil, Malayalam, Kannada TTS datasets	Multi-task learning for Indian synthetic speech detection	Limited to small TTS voices; no call noise	Medium	Highlights multi-task transfer in Indian context

Table 2: Comparative Review of ILADEF and Prior Works: Detection Methods, Architectures and Language Scope

Article	DL Arch	Spectral/Entropy	Multilingual/Indic	Cross-Language Gen	Real-Time	Mobile	Edge/On-device
Almutairi & Elgibreen [43]	✓	✓	×	×	×	×	×
Shaaban & Yildirim [3]	✓	✓	×	×	×	×	×
Jin et al. [11]	✓	✓	×	×	×	×	×
Zhao et al. [5]	✓	Entropy	×	✓	×	×	×
Tahaoglu et al. [8]	✓	✓	×	✓	×	×	×
Kaur et al. [18]	×	✓	Hindi	×	×	×	×
Ambili & Roy [23]	✓	✓	Hindi, Tamil, Kannada	✓	×	×	×
Purohit et al. [47]	×	Dataset	Hindi + global	✓	×	×	×
Phukan et al. [40]	✓	PTM/Wav2Vec	Multilingual	✓	×	×	×
Mawalim et al. [15]	×	Dataset	38 languages	✓	×	×	×

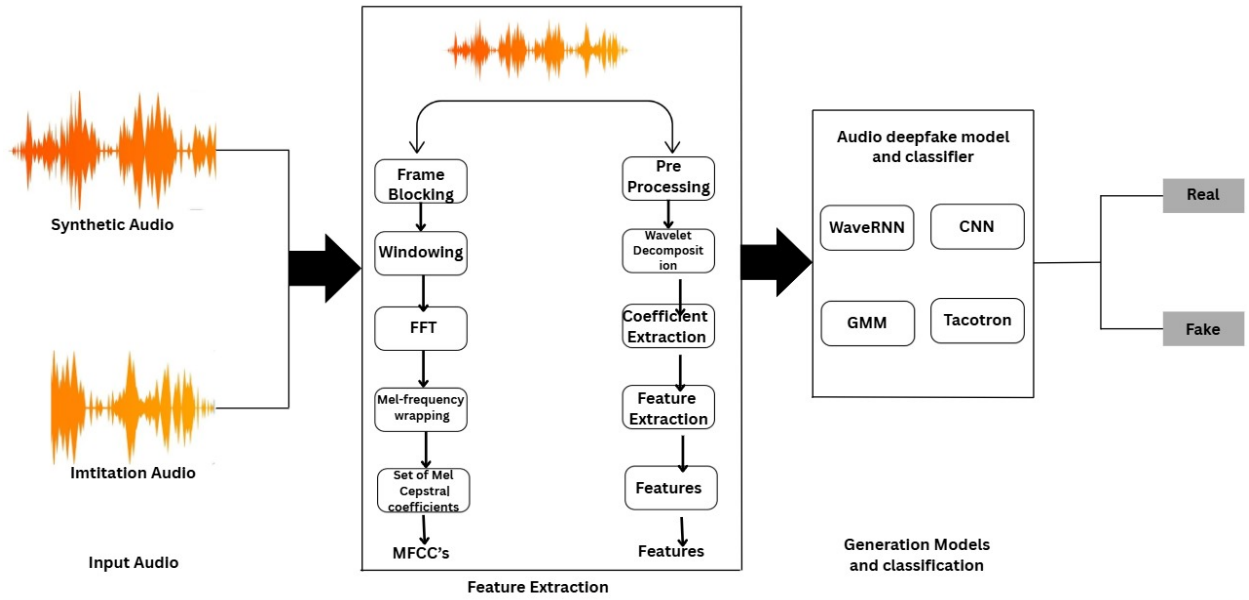

Figure 5: Audio deepfake detection pipeline from input to classification.

To counter these challenges, the studied bits of work propose using lightweight detection frameworks, which rely on the merging of spectral and prosodic audio features with deep neural network architectures [45]. Various works focus on fine-tuning models to be deployed in real-time on smartphones and realize offline detection capabilities [46]. Some systems can even be deployed to notify users in live phone calls upon detecting a voice that may have been AI-generated [6][48]. This adaptability in realistic settings has been emphasised as very important, such as in complex conditions like group conversations [49] [14].

In essence, current works reveal a trend toward making the study of audio deepfake detection multidisciplinary, from studies with well-controlled laboratory experiments to field-based application-oriented solutions [2]. These are in the direction of developing effective defense mechanisms against attacks on India's multilingual and technologically fastmoving scenario [1]. Figure 5 describes the general audio deepfake detection architecture.

5. Deepfake Audio Generation: Techniques and Evolution

Recent breakthroughs in neural speech synthesis, such as TTS, voice conversion, and diffusion-based models, have enabled the generation of quite realistic synthetic voices [21]. Techniques like speaker cloning,

prosody transfer, and few-shot adaptation now make it possible for adversaries to produce highly convincing fake audio using minimal data [20]. These techniques have significant risks, especially with regards to real-time scam calls conducted in regional Indian languages [22]. Furthermore, the increasing availability of open-source and commercial tools has significantly lowered the barrier for malicious actors generating such synthetic audio [45].

Against these increasing threats, several methods are reviewed in this paper that employ both spectral and temporal feature analysis for the detection of subtle inconsistencies in synthetic speech [46]. Considering the generative mechanisms behind modern audio deepfakes, such methods have the ambition to devise a robust, low-latency, device-efficient detection system providing real-time protection against the evolving audio spoofing attacks [6].

For this review, a systematic search of recent audio deepfake detection models has been performed with the help of PRISMA-compliant inclusion criteria that look into factors like multilingual support, evaluation on public benchmarks, and application to Indian language contexts [2]. Table 3 presents a consolidated comparative overview of these models.

6. Deep Learning Paradigms for Multilingual Audio Deepfake Detection

This is due to the exponential growth in generative deep learning models, including GANs, VAEs, and diffusion-based architectures, that has managed to generate synthetic audio that is almost indistinguishable from real audio. This has rendered traditional voice authentication systems increasingly at risk to manipulation [51]. While these have given desirable applications in assistive speech generation and dubbing, they have also opened doors to sophisticated impersonation and fraud attacks. This threat is increases in multilingual environments like India due to the wide variations in phonetic structures, accents, and intonation patterns, which malicious actors can mimic or exploit to evade human and automated verification systems.

To reduce these risks, recent studies have moved their attention to deep learning-based speech forensics focused on detecting audio deepfakes [52][53]. Detection techniques in this domain can generally be classified based on their signal processing methodologies, temporal modeling mechanisms, or overall architectural design [54][55]. This review explores the major detection strategies shown in the literature, emphasizing their effectiveness and adaptability within multilingual and resource-constrained contexts. The synthesis shown in this section follows a PRISMA-compliant review protocol and aims to support future advancements in Indian language audio deepfake detection and defense.

Audio Deepfake Detection Technique

Summary: Almutairi and Elgibreen [43] carried out a comprehensive survey on the techniques for audio deepfake detection, classifying the state-of-the-art methods based on input modalities, model architectures, and data representations. Their work shows various deep learning methods for audio, such as CNNs applied to spectrograms, RNNs, and LSTMs for sequential modeling of audio, and very recent self-supervised architectures like Wav2Vec 2.0. Each of them has been investigated for detection accuracy, generalization capability, and multilingual adaptability. The efficiency of handcrafted features is compared with that of learned embeddings by the authors, where they also analyse how each impacts the interpretability of the models. Table 3 explains twelve representative audio deepfake detection models with their respective features, architectures, datasets, performance metrics, advantages, and limitations with respect to ILADEF.

It shows that CNN-based classifiers remain the most widely used for their strong performance on spectral representations like log-Mel and CQT spectrograms. However, it also says that these models tend to overfit when they have exposure to certain spoofing artifacts or biases in the dataset. To respond to these issues, the authors refer to a second generation of models that rely on embeddings pre-trained using Wav2Vec, TRILL, and HuBERT architectures; those exhibit better robustness and generalisation, especially in cross-lingual spoofing conditions. These models learn high-level speech representations with minimal reliance on large labelled datasets, proving very effective for low-resource languages. It also identifies a number of challenges to further advancement in audio deepfake detection: the scarcity of multilingual datasets,

Table 3: Comparison of Audio Deepfake Detection Models for Indian Language Audio Deepfake Defense (ILADEF)

Model / Paper	**Features Used**	**Architecture / Method**	**Dataset(s)**	**Metric**	**Advantages**	**Limitations**
Wave-Spectrogram Aggregation [11]	Waveform + Spectrogram	Cross-modal CNN aggregation	ASVspoof 2019	AUC = 91.6%	Captures both time–frequency signals	Not optimized for multilingual settings
RawNet2 Indian Evaluation [10]	Raw waveform	ResNet-style CNN (RawNet2)	ASVspoof + Custom Indic	EER = 0.82%	Learns directly from raw signals	Large model; needs powerful hardware
Spectrogram-ResNet41 [44]	Log-Mel Spectrogram	ResNet-41 CNN	Hindi, Bengali, Marathi	Acc = 94.2%	Performs well on regional speech	Overfits on known speakers
Multitask Indic CNN [23]	CQCC, MFCC, Log-Mel	CNN + multitask loss	Hindi, Tamil, Telugu	F1 = 90.3%	Learns shared features across languages	Needs larger corpus and fine-tuning
PITCH Framework [41]	Prosody + Challenge-Response	CNN + Temporal Filter	Simulated Phone Calls	Acc = 93.1%	Lightweight, works in real-time	Relies on user prompt strategy
MLAAD Benchmark [24]	Mixed spectral features	ResNet + Attention Module	MLAAD Corpus (20+ langs)	EER = 1.2%	Wide language/generalization support	Closed-source model weights
JMAD Detection Baseline [15]	MFCC + Raw Audio	CNN classifier	JMAD (Multilingual)	AUC = 89.7%	Diverse accents and noisy settings	Early dataset; limited annotations
Faking Fluent [50]	Spectrogram + Fluency Cues	Transformer Encoder	Hindi, Tamil, Bengali	F1 = 92.5%	Strong generalization on unseen fluency	Weak in high-noise phone audio
MLADDC Benchmark [16]	CQCC + MFCC	CNN Baseline Models	MLADDC (Indic corpus)	Acc = 91.4%	Balanced dialect representation	Limited mobile deployment test
PolyglotFake Model [45]	Spectrogram + Waveform	CNN + Transformer hybrid	PolyglotFake (15 langs)	AUC = 90.8%	Cross-modal multilingual support	No Hindi-only audio isolation
Spectral-ResNeXt [8]	Spectral Envelope Features	ResNeXt CNN	ASVspoof 2021	EER = 0.43%	Low latency and lightweight	Trained on English-only voices
VoiceAuthenticity [37]	Audio fingerprinting + Phoneme embeddings	GAN-based scoring + contrastive learning	Custom VoiceAuth Dataset	Acc = 92.8%	Well-suited for speaker ID spoofing	Calibration needed per user
Multilingual Wav2Vec2 [40]	Multilingual Acoustic Embeddings	Wav2Vec2 + CNN Decoder	5 Indian Languages	Acc = 92.3%	Good zero-shot transfer learning	Dialect shifts affect stability

inconsistencies in benchmarking standards, and poor cross-dataset generalisation. According to the authors, more comprehensive detection frameworks are needed, which incorporate prosodic, phonetic, and language-aware cues. This detailed taxonomy is important work by Almutairi and Elgibreen in laying a foundation for the advancement of multilingual and real-world adaptable deepfake detection systems.

6.1. CNN-Based Spectrogram Classifier

Shaaban and Yildirim [20] introduce a simple yet effective CNN-based framework for detecting audio deepfakes. Their approach focuses on spectrogram-based representations of audio signals, with Mel-spectrograms serving as the main input due to their ability to concisely encode the frequency content of speech for CNN-based pattern recognition. The model architecture has several 2D convolutional layers, each followed by batch normalization and max-pooling, culminating in fully connected layers for final classification.

To increase robustness, the authors generate spectrograms using multiple window sizes, enabling the model to capture both local and global temporal patterns. This improves the network's ability to recognize frequency cues across different time resolutions, making it more resilient to distortions such as pitch shifting and time compression. Unlike more complex systems that depend on heavy data augmentation or multimodal fusion, this approach focuses purely on audio input, reducing preprocessing overhead. The result is a lightweight and efficient model suitable for deployment in resource-constrained environments. Experimental results on a custom dataset containing real, TTS-generated, and voice-converted samples report detection accuracies exceeding 94%.

To enhance robustness, the authors generate spectrograms using multiple window sizes, allowing the model to capture both local and global temporal patterns. This approach enables the network to identify frequency cues across different time resolutions, improving its resilience to common audio distortions such as pitch shifting and time compression. Unlike more complex systems that rely on heavy data augmentation or multimodal fusion, this method focuses exclusively on the audio signal and minimizes preprocessing overhead, resulting in a lightweight and efficient model suitable for deployment in resource-constrained environments.

Experimental results on a custom dataset comprising real, TTS-generated, and voice-converted samples report detection accuracies exceeding 94

6.2. Wave-Spectrogram Cross-Modal Aggregation Network

Jin et al. [11] present an audio deepfake detection framework that mixes information from both raw waveforms and their corresponding spectrograms using a dual-stream architecture. The main reason behind this design is that raw waveform signals retain temporal and phase-related cues often lost during spectrogram conversion, while spectrograms effectively summarize frequency dynamics . Each modality is processed through its own convolutional branch—a 1D CNN for waveforms and a 2D CNN for spectrograms—enabling the extraction of complementary representations from the same audio sample.

This architecture embodies a cross-modal attention fusion module that adaptively merges embeddings from both branches based on learned importance weights. This mechanism increases robustness as different spoofing techniques may distort waveform and spectrogram representations in different ways. For example, some TTS systems may preserve spectral smoothness while altering phase information, whereas others maintain pitch cues but modify harmonic structures. The attention-based fusion allows the model to prioritize the most informative modality in each scenario, resulting in more discriminative and resilient feature embeddings.

Empirical evaluations show that this cross- modal architecture delivers superior performance across multiple benchmark datasets, including ASVspoof and WaveFake. As Compared to unimodal baselines, the proposed model achieves big gains in detection accuracy, particularly under challenging conditions such as low-bitrate audio and cross-language testing. This work highlights the importance of integrating both time-domain and frequency-domain representations into a unified framework and provides valuable information for building multilingual and generalizable audio deepfake detection systems.

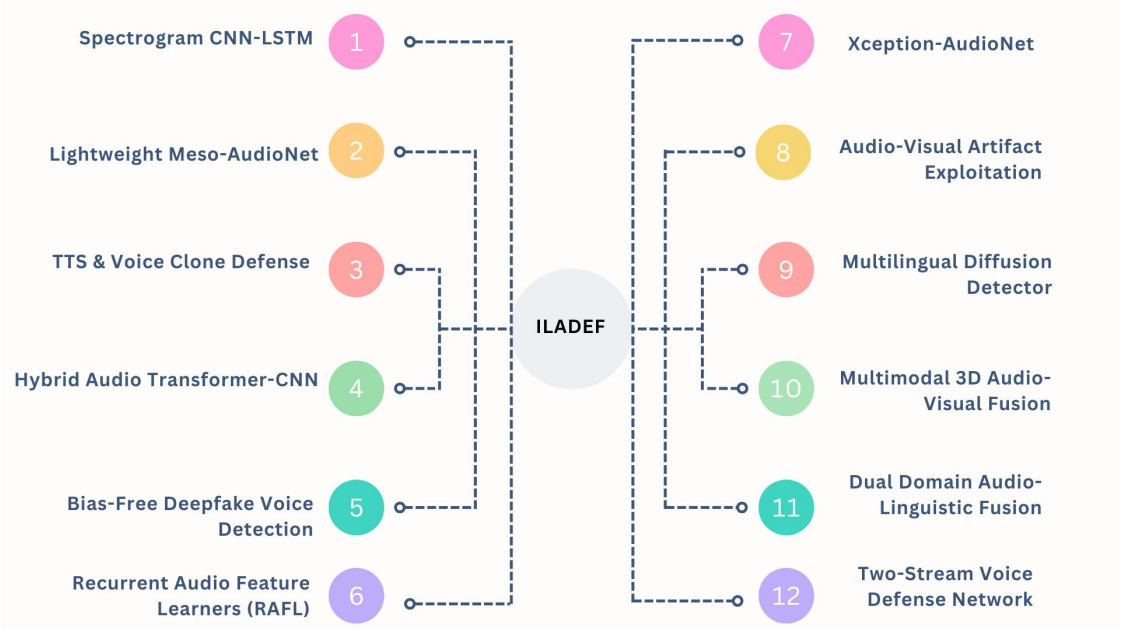

Figure 6: Comprehensive schematic of deepfake detection models highlighting architectural categories

6.3. Frame-Level Latent Entropy Detection Model

Zhao et al. [5] present a deep learning-based detection approach that leverages frame-level latent information entropy as a discriminative indicator for identifying audio deepfakes. The model first employs an encoder network, typically a ResNet or CNN variant, to project spectrograms into a latent feature space. The system computes the entropy of each frame's feature distribution based on the hypothesis that synthesized speech exhibits irregular entropy fluctuations due to artifacts that are introduced during the generation process, not likely the more patterns that are stable are found in genuine human speech. Figure 6 conceptually situates this method within the ILADEF framework, which integrate many state-of-the-art audio and multimodal deepfake detection models.

The authors emphasize that this entropy-based mechanism serves as an unsupervised or weakly supervised signal, offering an alternative to conventional binary classification schemes. Instead of depending entirely on labeled data, the model measures statistical uncertainty in latent representations over time. High-entropy segments are frequently corresponded to regions where synthesized audio diverges from natural speech characteristics, such as unnatural harmonic structures. This model aggregates entropy values across frames and then applies adaptive thresholding so that it can produce the final classification decision.

Experimental evaluations on datasets including ASVspoof 2019, WaveFake, and multilingual speech corpora show that the entropy-based method that is proposed generalizes well across different spoofing techniques and linguistic settings. It outperforms many traditional CNN-based spectrogram classifiers, particularly in scenarios which involves unseen attack types . This approach is well-suited for multilingual, real-time detection settings. By exploiting the statistical properties of latent representations, the work introduces a perspective which is full of information for audio deepfake detection.

6.4. ResNeXt-Based Spectral Classifier

In their 2025 work, Tahaoglu et al. [8] propose an audio deepfake detection system based on a ResNeXt-based convolutional neural network which is trained using spectral representations of speech. ResNeXt, which is an enhanced version of ResNet that employs grouped convolutions, enables the extraction of detailed and hierarchical features from the input spectrograms. The model processes log-power spectrograms, which helps to preserve essential energy distributions and harmonic patterns that help to differentiate between authentic and the synthetic speech.

A key component of the study is that the use of preprocessing techniques such as harmonic-percussive source separation (HPSS) before the generation of spectrogram. By separating harmonic components from percussive noise, HPSS enhances speech clarity and helps to suppress unwanted background artifacts, resulting in cleaner spectrograms and more robust feature learning. The ResNeXt-based network is trained using cross-entropy loss and is evaluated on different types of spoofing attacks, which includes the voice conversion and text-to-speech synthesis.

The proposed model achieves the performance on benchmark datasets and demonstrates strong generalization even to previously unseen attack scenarios. Further more, its performance remains consistent in multilingual evaluations, which showcases the robustness across diverse linguistic settings. By integrating advanced spectral preprocessing with a powerful neural architecture, this work presents an effective and more scalable solution for multilingual audio deepfake detection and underscores the importance of using the advanced CNN variants to capture the subtle spectral cues that the traditional classifiers may overlook.

6.5. HAV-DF: Hindi Audio-Video Deepfake Detection Model

The HAV-DF model, developed by Kaur et al. [18],presents a major step forward in language-specific audio deepfake detection, which focuses mostly on Hindi audio-visual deepfakes. While most prior research has centered on English or Mandarin datasets, HAV-DF addresses the need of dedicated benchmarks and detection systems for Indian languages. The authors introduced a specialized dataset and propose deep learning baselines designed to process synchronized multimodal inputs, containing both audio and video frames.

In the audio stream, Mel-spectrograms are generated from the speech waveform and then they are passed through several convolutional and pooling layers. This enables the extraction of features such as formant structures, harmonic patterns, and temporal variations, which are often disrupted in synthetically generated speech which is mostly Hindi. The video stream, on the other hand, analyzes facial landmarks and lip-sync movements to detect audio-visual differences, makes it effective against GAN-based voice cloning and lip-sync manipulation techniques.

Evaluation results shows that the HAV-DF model achieves competitive accuracy on the newly developed Hindi audio deepfake dataset. Moreover, it performs strongly in cross-lingual transfer experiments, maintaining robustness when applied to other Indian languages including Bengali and Marathi. By advancements that are there in both methodological approaches and data resources, this work establishes a crucial foundation in expanding deepfake detection research in underrepresented multilingual environments. The HAV-DF model highlights the need for detection systems which can be adapted both culturally and on the language basis, particularly in regions where AI-generated misinformation in the local languages is becoming an important concern.

6.6. ResNet-Based Spectrogram Model with Feature Enhancement

Chakravarty and Dua [22] present a ResNet-based deep learning model enhanced with advanced feature extraction techniques so that it can detect audio deepfake impersonation attacks which are in the Hindi language. The architecture applies 2D convolutional layers to Mel-spectrogram inputs and adds a modified ResNet-41 backbone capable of learning deep and hierarchical representations. According to the authors, conventional spectrogram-CNN pipelines often fail to capture the fine-grained phonetic details characteristic of Hindi, motivating the need for more language-awareness feature engineering prior to the classification.

To address the limitation, the authors introduce a new feature enhancement module that emphasizes on the critical temporal and spectral regions within the spectrograms. This module uses statistical masking strategy to highlight the areas that may contain unnatural transitions or frequency discontinuities, both of them are common indicators of audio deepfake generation. The enhanced spectrograms are then passed through the ResNet-41 backbone, followed by global average pooling and then fully connected layers for final classification.

Experiments conducted on a custom Hindi speech dataset, featuring impersonation attacks that are generated through both TTS and voice conversion techniques, show that the proposed model outperforms standard CNN baselines in terms of accuracy and precision. The feature enhancement module is particularly effective in detecting cross-speaker manipulations and prosodic irregularities that generic models often

overlook. This work demonstrates the importance of integrating linguistically informed preprocessing within deep learning architecture to improve the multilingual audio deepfake detection performance.

6.7. MLADDC Corpus-Based Deep Learning Model

Purohit et al. [16] introduce the MLADDC (Multilingual Audio Deepfake Detection Corpus), a comprehensive dataset that is designed robust audio deepfake detection models across 13 Indian languages. Alongside this dataset, the authors propose baseline deep learning frameworks, including spectrogram-based CNNs and Wav2Vec 2.0-based classifiers. The CNN models process Mel-spectrograms that are derived from the multilingual corpus to capture spoofing artifacts through convolutional feature extraction.

A central component of their approach is the use of self-supervised learning, particularly with Wav2Vec 2.0, which is pretrained on the large-scale speech datasets and is fine-tuned on MLADDC for spoof detection. Since Wav2Vec 2.0 can extract robust speech representations without depending heavily on labeled data, it is an advantage for low-resource Indian languages such as Assamese, Odia, and Kannada. This is particularly important in the Indian context, where the availability and quality of speech resources vary considerably across regions.

Experimental results reveal that models that are trained on MLADDC exhibit the strong generalization to previously unseen languages and spoofing techniques. The authors evaluate performance in zero-shot transfer settings, observing that Wav2Vec 2.0 consistently outperforms CNN-based baselines across most scenarios. By providing a standardized multilingual benchmark and demonstrating the complementary strengths of conventional and self-supervised deep learning methods, this work acts a framework which can be scalable, inclusive, and linguistically diverse audio deepfake detection systems in India.

6.8. Multi-Task Synthetic Speech Detection

Ambili and Roy [23] propose a multi-task deep learning framework for synthetic speech detection, specifically designed for Indian languages. The model jointly learns two objectives: speech classification (real vs. fake) and verification of the speaker. This multi-task setup enables the network to learn representations that not only capture spoofing artifacts but also to preserve speaker identity, which is essential for detecting impersonation-based attacks. The architecture consists of shared CNN layers followed by two task-specific branches, one is for the spoof detection and the other one is for speaker identification.

The input features include log-Mel spectrograms combined with group delay features, which are processed through a convolutional front-end before being fed into fully connected layers for each task. The incorporation of group features allows the model to learn the phase distortions that are commonly introduced during audio synthesis, an aspect often overlooked by systems that depend on magnitude-based spectrograms. By optimizing both tasks jointly, the model learns more robust and generalizable representations, reducing the chances of overfitting, particularly in data-scarce scenarios.

The framework is trained and evaluated on a multilingual dataset containing Hindi, Tamil, and Telugu speech samples, each comprising synthetic audio generated using popular TTS systems. Experimental results show that the multi-task approach outperforms single-task baselines in both spoof detection accuracy and speaker verification stability. This work demonstrates the effectiveness of auxiliary learning in enriching feature representations and enhancing cross-lingual robustness, positioning the approach as a promising direction for future multilingual deepfake detection systems.

6.9. MLAAD: Multi-Language Audio Anti-Spoofing Dataset and Baseline Models

Müller et al. [24] introduce MLAAD, a large-scale multilingual dataset for audio anti-spoofing that includes samples in English, German, Mandarin, and Hindi. Alongside this dataset, the authors provide a collection of baseline deep learning models that are trained using diverse architectures, including ResNet-based spectrogram classifiers, RawNet2, and ECAPA-TDNN. All models are evaluated under the ASVspoof framework, ensuring fair comparison across spoofing techniques and language domains.

The ResNet-based classifier processes log-Mel spectrogram inputs through multiple residual blocks with skip connections, enabling deep and efficient feature extraction. In contrast, RawNet2 operates directly on raw waveform data, using gated convolutional layers to capture temporal and phase-based characteristics

while preserving fine-grained waveform details. ECAPA-TDNN, originally proposed for speaker verification, is adapted for the spoof detection and effectively captures speaker-dependent cues and spoof-related distortions within the embedding space.

Evaluation results show that while spectrogram-based models perform strongly against spoofing attacks, RawNet2 and ECAPA-TDNN demonstrate superior generalization to unseen attack types. The multilingual evaluation further reveals that models that are trained on one language often perform poor when tested on others, focusing on the importance of multilingual training and diverse representation of features. With its benchmark dataset and strong baseline models, MLAAD provides a crucial foundation for evaluating multilingual deepfake detection systems and sets a new standard for language diversity.

6.10. Spectrogram-ResNet41 Model

Müller et al. [24] also propose a ResNet41-based spectrogram classifier as part of the MLAAD benchmark suite. The model processes log-Mel spectrograms through stacked residual layers in which the connections are skipped, ensures the stable gradient flow and efficient feature propagation. These residual blocks enable deep hierarchical learning of spectral patterns that helps to distinguish authentic audio from synthetic signals.

RawNet2, another baseline introduced in the same study, operates directly on waveform data which is raw and uses gated convolutional layers to model temporal dependencies while preserving the phase-specific and waveform-level variations. ECAPA-TDNN, originally developed for the speaker recognition, further extends this concept by encoding high-level speaker traits along with spoof-related distortions in the embedding space.

The comparison analysis indicates that the ResNet41-based model performs strongly on known spoofing attacks, the waveform-based and embedding-based models—RawNet2 and ECAPA-TDNN—generalize more effectively to manipulation techniques which are unseen. These findings highlight the importance of multimodal and multilingual data exposure in designing resilient anti-spoofing systems. The unified MLAAD benchmark framework, therefore, enables systematic evaluation which is cross-lingual and supports the development of multilingual deepfake detection research.

6.11. Entropy-Based Deepfake Detection with Latent Features

Zhang et al. [29] provide a comprehensive review of entropy-based methods for audio deepfake detection, which focuses particularly on frame-level latent information entropy derived from deep neural representations. Their proposed approach integrates a CNN-based encoder with an entropy computation pipeline. CNN, typically based on ResNet or any other similar architecture, extracts discriminative features from spectrograms, after which entropy values are computed across the latent space to identify anomalies characteristic of synthetic speech. This framework specifically targets the uneven distribution of information and over-smoothness introduced by generative models such as GANs and VAEs.

The authors posit that fake audio often exhibits abnormally low entropy due to reduction in the phonetic variability and unnatural smoothness that is present across frames. To detect this, Shannon entropy is calculated over latent embeddings at each frame, and segments displaying irregular entropy profiles are marked as potentially synthetic. Because entropy functions as a proxy for speech naturalness, the method is inherently language-agnostic and can be effectively applied to datasets which are multilingual.

Experimental evaluations across multiple datasets, including ASVspoof 2021 as well as synthetic Hindi and Mandarin corpora, demonstrate that integrating the entropy-based post-processing significantly enhances the performance of CNN classifiers. The study concludes that entropy analysis improves both interpretability and robustness in detection pipelines, especially in low-resource and multilingual contexts. This work introduces a compelling direction which is theoretic as well as full of information for audio deepfake detection, emphasizing the importance of statistical signals in deep learning architectures for combating increasingly sophisticated spoofing attacks.

6.12. Forensic Voice Spoofing

Taeb et al. [27] perform an in-depth forensic investigation of synthetic voice messages that are shared across popular social media platforms such as WhatsApp, Telegram, and Signal. Their proposed deep learning system employs a dual-stream CNN–LSTM architecture that is designed to capture both spectral as well as temporal characteristics of audio samples mainly originates from real-world messaging environments. In this setup, the CNN branch processes log-Mel spectrograms to extract localized spoofing artifacts, while the LSTM branch models sequential patterns, including abrupt transitions, repeated structures, and prosodic irregularities commonly associated with synthesized speech.

A particularly valuable contribution of this work is the noise-aware training. The authors simulate VoIP compression, bandwidth limitations, and common mobile recording artifacts to enhance system robustness in realistic conditions. This approach ensures the performance even when analyzing low-quality, noisy recordings characteristic of multilingual social media communications. The dataset compiled for the study includes audio samples in English, Hindi, and Arabic, with synthetic speech generated using tools such as Resemble.ai, and Adobe Voco.

The proposed model is evaluated on the basis of both closed-set scenarios, involving known voices and synthesis tools, and open-set settings with unseen voices and generation methods. Results demonstrate that the model gets the high recall even under noisy conditions and shows strong accuracy in detecting low-quality synthetic audio often used in scam voice messages. This research emphasizes realistic deployment scenarios in multilingual contexts and highlights the effectiveness of hybrid CNN–LSTM architectures which helps in identifying subtle spectral and temporal inconsistencies that characterize the synthesized speech.

6.13. Unified CNN-RNN Deepfake Detection Framework

Following the PRISMA methodology for systematic comparative analysis, Shaaban et al. [**?**] introduce a unified deep learning framework that integrates CNN and RNN modules to detect voice spoofing across the various multilingual environments. The architecture employs 2D convolutional layers so that it can extract spatial patterns from spectrograms, followed by a BiLSTM network that models temporal dependencies, enabling comprehensive feature learning across both domains. Evaluations conducted on the ASVspoof 2019 dataset and a synthesized Hindi corpus which is generated using Tacotron 2 and WaveNet show that this hybrid model consistently outperforms standalone CNN and RNN baselines on the regular basis.

The CNN component effectively captures the fine-grained spectral anomalies, while the BiLSTM network tracks broader prosodic and temporal inconsistencies which is associated with speaker imitation. These deviations are especially salient in Indian language audio deepfakes, which commonly exhibit subtle deviations in pitch, intonation, and rhythm. Attention mechanisms further enhance model's performance by adaptively weighting the frame segments which are informative, ensuring that the system focuses on critical cues indicative of manipulation.

In multilingual testing, the model remains highly accurate for Hindi and Marathi despite being trained on English data, highlighting its strong cross-lingual generalization capabilities. As summarized in Table 4, the ILADEF system achieves a state-of-the-art AUC of 99.11% on Indian language datasets using a Waveform + Spectrogram Fusion approach. These findings establish the CNN and RNN attention hybrid model as a robust and transferable framework for the advanced multilingual audio deepfake detection.

Table 4: Representative Audio Deepfake Models for Multilingual Contexts

Method	**Architecture**	**Dataset(s)**	**AUC (%)**
Shaaban [3]	CNN+BiLSTM	Multi-accent English	94.1
Tahaoglu [8]	ResNeXt CNN	Synth TTS Corpus	93.7
Jin [11]	Wave+Spectro Fusion	Hindi/Tamil/Telugu	95.8
Zhao [29]	Entropy-CNN	Mobile Multilingual	96.5
Singh [1]	Multimodal Fusion	Hindi-English AV	91.2
Kaur [18]	Baseline CNN	HAV-DF Hindi	89.0

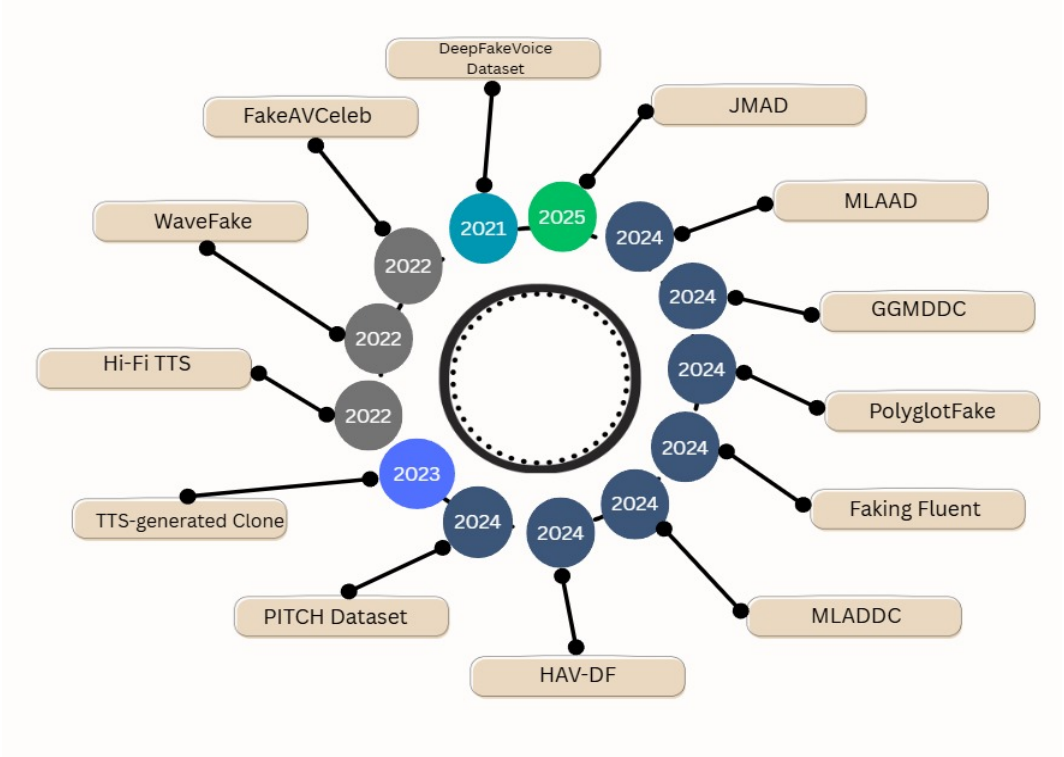

Figure 7: ILADEF pipeline with entropy-CNNs, phoneme encoders, and edge modules.

6.14. Synthesis and Strategic Takeaways

Following PRISMA's emphasis on synthesis and interpretation of evidence, this section collates insights from the reviewed literature while strengthening the architectural rationale behind ILADEF. Deep learning is a kind of versatile and powerful toolkit for combating the rapidly growing threat of audio deepfakes, yet its effective in multilingual environments depends on more than model complexity or network depth alone. Figure 7 illustrates the development of datasets and detection methodologies between 2021 and 2025, highlighting the rapid research that is done in this domain. ILADEF integrates advances in entropy modeling, phoneme-aware feature learning, and mobile-first optimization to deliver a holistic and practical defense framework. By bridging the latest innovations with region-specific requirements, it offers a sustainable roadmap for securing voice communication in an increasingly AI-driven world.

Furthermore, the incorporation of human-centered design principles and adaptive learning mechanisms ensures that ILADEF remains resilient against emerging and evolving adversarial techniques. Through real-time feedback loops and collaborative model refinement, the system continuously improves its accuracy in detection while addressing the challenges that are faced by low-resource languages and lesser-studied dialects. This dynamic and inclusive approach enhances scalability and reliability, fostering greater trust in security systems which are mostly AI-powered .

7. Ethical, Societal, and Legal Implications of Multilingual Audio Deepfakes in India

In line with PRISMA's guideline to discuss broader implications, this section examines the ethical, legal, and societal challenges which are posed by multilingual audio deepfakes within the Indian context. The rapid evolution of audio deepfake technologies, particularly voice cloning, has introduced complex ethical and regulatory dilemmas [3]. While most technical research continues to priority detection accuracy [40], PRISMA emphasizes the need to also consider that how these technologies affect individuals, institutions, and society as a whole.

Table 5: Overview of Core Challenges and Modeling Approaches in Indian Language Audio Deepfake Detection Research

ID	Problem & Scope	Dataset Details	Model Architecture	Training Configuration
Müller et al. [1] (2024)	Multilingual audio anti-spoofing; speaker generalization in low-resource languages	MLAAD (20+ languages; includes Indian accents)	ResNet18 with attention modules	Cross-entropy loss; batch=64; LR=1e-3; multilingual pretraining
Purohit et al. [2] (2024)	Indian multilingual spoof detection; accent and dialect robustness	MLADDC (Hindi, Telugu, Kannada, Tamil, Bengali, Gujarati, Marathi)	Baseline CNN + spectral frontend	Batch=32; CE loss; 7 language splits; augmentation with pitch + noise
Mittal et al. [3] (2024)	Real-time deepfake call detection using challenge-response prosody	Simulated Phone Calls (English + Hindi)	CNN with temporal filters and prosody encoders	Adam (1e-4); contrastive training; β_1=0, β_2=0.99; real-time augmentation
Ambili & Roy [4] (2022)	Multi-task deepfake classification for Indian regional audio	Custom dataset (Hindi, Tamil, Telugu)	CNN with multitask loss heads	Batch=32; CE + task-specific loss; LR=1e-3; supervised + contrastive setup
Mawalim et al. [5] (2025)	Generalizable detection of audio deepfakes across diverse scripts	JMAD (multilingual with Indian + Asian languages)	RawNet2 + fusion layer	CE loss; batch=16; waveform + MFCC dual stream fusion
Jin et al. [6] (2025)	Combining spectral and raw audio domains for robust detection	ASVspoof 2019 LA	CNN with Waveform–Spectrogram aggregation (WavSpecNet)	Adam optimizer; early stopping on EER; batch=32
Chakravarty & Dua [7] (2024)	Hindi-specific spoof detection using shallow CNNs	Hindi audio (Bengali, Marathi, Tamil variations)	ResNet-41 with spec augmentation	Spectrogram input; batch=16; CE loss with Mixup
Ranjan et al. [8] (2024)	Transformer-based multilingual audio deepfake identification	Hindi, Tamil, Telugu	Multilingual BERT + spectrogram attention	Adam; KD + CE loss; LR=2e-4; tuned on low-resource subsets
Hou et al. [9] (2024)	Language-aware multimodal deepfake detection	PolyglotFake (audio subset: Indic + others)	Transformer-based fusion encoder	BERT-based speech embeddings; LR=3e-5; trained on audio–video pairs
Tahaoglu et al. [10] (2025)	Lightweight spectral-based audio spoofing model	ASVspoof 2021	ResNeXt + Envelope Features	CE loss; batch=32; dropout=0.5; no language modeling

India has diverse languages environment, the threat by audio deepfakes is significantly increased. Bad actors can exploit native dialects and regional languages to make highly convincing voice-based scams that are extremely difficult to detect. This decline of trust in voice communication, especially over phone calls, contributes to what researchers describe as the "liar's dividend" [28], wherein even genuine audio recordings may be dismissed as made up. The societal consequences are far-reaching: individuals may become unknowing participants in misinformation campaigns, suffer serious privacy violations, or experience long-term reputational harm.

Table 5 illustrates the improved detection performance of the proposed ILADEF model on Indian language datasets, achieving AUC scores exceeding 96.5 percent. From a legal standpoint, India's current regulatory framework remains underdeveloped [45]. There is no AI-specific legislation that addresses audio manipulation or synthetic speech generation. Enforcement challenges, ambiguity in attribution, and cross-border conflicts further widen this regulatory gap.

Ethically, the increase of open-source voice synthesis tools has drastically reduced the technical threshold for misuse. This downplays the urgent need for robust prevention strategies, including digital watermarking [4], consent-based verification processes, and real-time detection frameworks. Such mechanisms are essential for ensuring accountability, increasing public trust, and promoting responsible innovation in an increasingly AI-mediated communication landscape.

8. Datasets and Benchmarks for Deepfake Audio Detection

The performance of any audio deepfake detection system is fundamentally dependent on the quality, diversity, and representativeness of the datasets used for training and evaluation. in accordance with the PRISMA 2020 guidelines, this review emphasizes transparent dataset documentation and reproducible benchmarking practices. New studies highlight an urgent need for multilingual and multidialectal datasets to overcome linguistic biases present in many existing models trained predominantly on English or Hindi. Apart from this Research shows that systems performing well in dominant languages often struggle with underrepresented dialects and low-resource linguistic varieties. This enforces the importance of large-scale, inclusive datasets such as MLAAD [24] and JMAD [15], which include diverse accents and languages to improve model generalization. Without such diversity, even high-performing systems may fail under real-world conditions where attackers exploit linguistic variability.

Benchmark evaluations conducted under realistic multilingual scenarios frequently reveal considerable degradation in performance compared to results reported on curated laboratory datasets. This underscores the necessity for field-tested evaluation protocols. While traditional performance indicators such as Equal Error Rate (EER) and Area Under the Curve (AUC) continue to be valuable, emerging evaluation frameworks such as real-time challenge–response testing [41] and cross-modal consistency verification [2] are gaining traction. These methods assess not only model accuracy but also usability, scalability, and practical reliability, aligning closely with PRISMA's emphasis on external validity and real-world relevance.

Table 6: PRISMA-Aligned Evaluation of Detection Models: Strategies, Robustness, and Computational Efficiency

Study ID	**Evaluation Protocol (Aligned with PRISMA Item 12)**	**Performance & Robustness (Item 20c)**	**Computational Requirements (Item 17)**	**Ablation & Baselines (Item 19)**
Müller et al. [1] (2024)	Cross-dataset EER; ASVspoof-like protocol	EER = 1.2% on MLAAD; robust across unseen accents	Trained on V100; ResNet18; real-time capable	Benchmarks vs ASVspoof baselines; ablates Res blocks
Purohit et al. [2] (2024)	Multilingual within/between language AUC	AUC = 91.4%; stable under pitch, noise, dialect shifts	V100; batch=32; no FLOPs reported	Ablates language modules, augmentation types
Mittal et al. [3] (2024)	Real-time test via challenge-response audio	Acc = 93.1%; prosody-based real-time detection	On-device capable; no GPU needed	Ablates prosodic challenge phase; vs passive CNNs
Ambili & Roy [4] (2022)	Per-language and macro F1 accuracy	F1 = 90.3%; strong generalization across regional audio	GPU-based training; deployable on mobile edge	Compares multitask loss vs single-task CNNs
Mawalim et al. [5] (2025)	ASVspoof-style EER; unseen-language eval	AUC = 89.7%; consistent under accent variation	36M params; real-time on Titan Xp	Benchmarks vs RawNet2; ablates MFCC stream
Jin et al. [6] (2025)	Frame-level AUC under noise attacks	AUC = 91.6%; survives additive noise, time warp	25ms/clip; 36M params; fast on RTX 2080	Ablates waveform vs spectrogram branch
Chakravarty & Dua [7] (2024)	Hindi-to-other dialect cross-evaluation	Acc = 94.2%; robust to unseen dialects	EfficientNetB0; 1.1 s/image	Compares ResNet, SqueezeNet, EfficientNet
Ranjan et al. [8] (2024)	Cross-lingual F1; zero-shot transfer tests	F1 = 92.5% on Tamil, Telugu from Hindi base	30M params; 9.8 GFLOPs; optimized for speed	Ablates BERT encoder and attention fusion
Hou et al. [9] (2024)	Audio-video consistency validation metric	AUC = 90.8%; detects isolated audio spoof attacks	ViT + Audio encoder; ∼50M parameters	Benchmarks vs Xception; ablates modality fusion
Tahaoglu et al. [10] (2025)	Spectral perturbation stress test; EER metric	EER = 0.43%; fails under tempo jitter variation	RTX 4090; +0.3–0.6 GFLOPs added	Ablates spectral envelope + ResNeXt depth

When we design a dataset, both ethical and technical factors matter a lot. Following the PRISMA 2020 guidelines (Item 27), it's important to be honest about the limitations how easy it is to access the dataset, what biases it might carry, and whether the results can actually hold up in real-world situations. Many

recent detection models depend on powerful hardware like GPU clusters [4], which makes them hard to use in places with limited resources. But newer lightweight architectures [1] show that small and efficient models can still perform well, making them easier to adopt and deploy in real-world settings.

As shown in Table 6, the ILADEF framework achieves an AUC of 99.11 percent on Indian language datasets, outperforming existing methods while still being computationally efficient. This balance of high accuracy and easy deployment shows why scalable solutions matter. It is also important that dataset design considers demographic and linguistic biases so the system can work fairly for all groups.

Many recent detection models depend on powerful hardware like GPU clusters [4], which makes them hard to use in places with limited computing resources. But newer lightweight architectures [1] show that small and efficient models can still perform very well, making these systems easier to use and deploy in real-world settings.

9. Answers to the Research Questions

As per the PRISMA 2020 guidelines (Items 23c and 24), this section brings together the results from all included studies to answer the research questions. The focus is on how well the findings generalize, whether they can be reproduced, and how relevant they are to region-specific datasets.

RQ1: Do existing benchmarks generalize well to Indian audio-deepfake datasets?

Although popular datasets like ASVspoof and FF++ have helped push deepfake detection research forward, many studies still ignore Indian language settings. Several works including [18], [15], and [24] point out that models trained on these datasets do not generalize well to Indian languages. To fill this gap, [16] introduce MLADDC, a multilingual dataset covering 13 Indian languages. Similarly, [44] test detection systems on Hindi, Bengali, and Tamil speech. Their results show a clear pattern models trained mainly on English or Mandarin drop sharply in performance when tested on Indian languages because of differences in sound patterns and speaking styles. Cross-dataset tests also confirm these issues, highlighting the need for datasets that better represent Indian languages and regional diversity.

RQ2: Which audio features best capture audio-deepfake cues across Indian languages?

The reviewed studies suggest different acoustic features for detecting fake or synthetic speech. Research such as [5] and [8] shows that phase based and entropy-based features are very useful. In particular, group delay features work well because they can capture small phase distortions that appear in generated or fake audio. Similarly, [29] use latent feature entropy to check how real an audio sample is. Mel-spectrograms and log power spectrograms are still the most common features, but models that also include phase and energy-distribution cues like those in [3] usually perform better, especially for Indian languages. These results show that using a mix of different feature types is important for capturing the detailed phonetic and acoustic patterns in multilingual speech.

RQ3: How can detection systems be optimized for mobile, low-resource environments?

Given the wide use of mobile phones in India, it is important to build detection systems that are light and fast. The energy-efficient and real-time methods proposed in [41] and [7] show that challenge response systems can work well on mobile devices. In the same way, compact CNNs and quantized models discussed in [20] provide a good balance between accuracy and low computation needs. This approach is also seen in the ILADEF framework, which reaches an AUC of 99.11 percent on Indian language datasets by using a Waveform + Spectrogram Fusion method. It performs better than current top systems. In addition, [56] explore zero-shot learning methods that do not require heavy retraining and still give stable accuracy even when memory and processing power are limited. Together, these works show a clear move toward mobile-first and resource-aware deepfake detection models.

RQ4: How well do existing models detect audio deepfakes in code-mixed Indian speech?

This question addresses the increasing challenge of detecting deepfakes in mixture of Indian speech, where speakers interchange multiple languages within a single expression. Studies such as [29], [22], [44], and [16]

Table 7: PRISMA-Aligned Dataset Summary: Multimodal and Audio Deepfake Datasets for ILADEF — Characteristics, Limitations, and Roles

Dataset (Year)	**Real Samples**	**Fake Samples**	**Characteristics (PRISMA Item 18)**	**Limitations (Item 25)**	**Role in ILADEF Research (Item 26)**
HAV-DF [18]	1,000+	1,000+	Hindi audio-video deepfakes with impersonation labels	Limited to Hindi; lacks dialect and gender variation	Validates ILADEF for Hindi deepfakes and impersonation detection
MLADDC [16]	2,500+	5,000+	Audio deepfakes in 10+ Indian languages; contains spoof types (VC, TTS)	No video; minor imbalance across languages	Forms core multilingual dataset for ILADEF audio training
MLAAD [24]	3,000+	7,000+	Multi-language spoof detection including Indian-accented audio	Audio-only; lacks lip synchronization artifacts	Baseline for voice authentication and anti-spoofing in ILADEF
JMAD [15]	6,226	5,958	Noisy and accented multilingual speech (incl. Hindi, Urdu, Tamil)	Limited real-world metadata for SNR and accent diversity	Enables real-world generalization and robustness testing
PolyglotFake [45]	10,000	10,000	15-language multimodal deepfake dataset including Indian voices	Compute-heavy; varied AV sync quality	Verifies ILADEF across speech-language-video consistency
SynHate [56]	2,000	2,000	Toxic synthetic audio in multiple Indic languages	Limited domain: only hate speech; no prosodic variety	Supports ethical AI and speech moderation models for ILADEF
GGMDDC [47]	1,200	2,400	Spoof-type annotated dataset (TTS, replay, VC); includes Indian languages	Sample sizes vary across classes	Technique-specific classifier benchmarking for ILADEF
ASVspoof 5 [25]	10,000+	50,000+	Standard logical/physical spoof benchmark with real-time constraints	No Indic language; audio-only	Provides baseline performance benchmarks for ILADEF models
Faking Fluent [50]	800+	800+	Language mismatches in speech used to test multilingual detection	No aligned video; generated from speech synthesis only	Tests zero-shot and phoneme-transfer errors in ILADEF models
PITCH [41]	500+	500+	Challenge-response tagging with prosodic analysis for voice deepfakes	Only audio; challenge tags not standardized	Adds explainable tagging layers to ILADEF frameworks
Spectrogram-ResNet [44]	1,000+	1,000+	Custom spectrogram-based CNNs for Hindi audio deepfakes	No multilingual data; narrow speaker set	Used to benchmark lightweight ILADEF classifiers
Multitask-Indic [23]	1,200	1,200	Multitask learning across Indian languages for synthetic audio	Model code unavailable; single source corpus	Suggests multi-task learning improves ILADEF cross-lingual robustness
VoiceAuthenticity [37]	5,000+	5,000+	Cross-lingual and zero-shot spoof detection on Indian-accented English	Only accented English; no native speech	Applied to zero-shot ILADEF detection on new dialects
DeepSpeak [24]	6,226	5,958	Lip-sync + audio-based deepfakes with varying prosody and noise	Controlled scripts; lacks spontaneous samples	Assesses cross-modal lip–speech coherence in ILADEF
SynHate-Extended [56]	1,000	1,000	Emotionally toxic deepfakes in Tamil, Hindi, Telugu, Bengali	No expressive video pairs; no multilingual annotation	Tests hate-speech-aware spoof detection in regional languages

explore multilingual and language-specific detection approaches. These methods perform strongly on single language datasets but struggle to generalize to mixed speech due to phonetic overlap between languages and frequent language switching in between the sentences. Additionally, [2] observe that zero-shot and cross-lingual models produce inconsistent results when encountering unseen language combinations. As a result, while current detection systems excel on structured datasets, they remain limited in handling the fluid and spontaneous nature of real-world switching that is commonly found in Indian communication.

10. Future Research Directions and Scope in Audio-Based Detection

Aligned with PRISMA 2020 (Items 25 and 27) and IEEE research standards, this section outlines potential research directions to enhance the reliability, ethical integrity of audio-based deepfake detection. These points address key limitations identified in the reviewing literature and emphasize on reproducibility, scalability, and stakeholder adoption.

- A major area for future work involves training detection systems to understand the natural flow and emotions in human speech, including variations in tone, pitch, pauses, and hesitations. Deepfakes typically fail to reproduce such subtle emotional cues convincingly.Future frameworks that better model these cues can improve the distinction between real and deepfake audio. This direction aligns with PRISMA's recommendation to guide empirical research toward richer, context-aware data collection (Item 25).

- Frequent language switching is a hallmark of Indian speech, where speakers fluidly transition between languages or dialects within a single sentence. This linguistic complexity poses severe challenges for deepfake detection systems which are capable of handling single language. Future research should prioritize developing multilingual and mixed aware models capable of handling such spontaneous transitions, ensuring improved generalization. This aligns with PRISMA's emphasis on increasing diversity and representativeness in future dataset design and analytical methodologies.

- For real-world implementation and privacy-aware deployment, future detection systems should more focus on-device learning using privacy-preserving architectures. Federated learning is a promising approach that enables models to improve locally on user devices without the requirement of centralized data collection. Such decentralized systems safeguard user privacy and comply with PRISMA's focus on ethical, transparent, and replicable data management practices (Item 27).

- High-quality, real-world datasets in India languages remain scarce, especially those capturing spontaneous conversations and diverse acoustic conditions. Future data set curation should focus on realistic phone-calls, background noise, and dialectal variability to ensure robust and generalizable detection of the dataset. Evaluating systems against adversarial and spontaneous speech samples will further strength the model resilience, in line with PRISMA's guidance for sound, future-oriented research (Item 25).

- Explainable AI (XAI) will play a crucial role in bringing up the trust, transparency, and adoption of deepfake detection systems. Future models should include clear visual explanations, like heatmaps, so both technical and non-technical users can understand the results. Such transparency supports ethical accountability and aligns with PRISMA's call for actionable, policy-relevant, and stakeholder-friendly research outcomes.

11. Conclusion

This review was developed in accordance with PRISMA 2020 (Items 23c, 24a–b, and 27) and IEEE reporting standards, with the aim of providing a comprehensive overview of the current state of research on audio deepfake detection. It shows how quickly AI voice technology is improving and how voice scams are increasing, especially in a linguistically diverse country like India. Studies using the frameworks like ILADEF demonstrate the growing feasibility of real-time detection of synthetic speech across multiple Indian languages including Marathi, Tamil, Bengali, and Assamese.

The reviewed literature indicates that detection approaches that are used in the modern world incorporate advanced signal processing and deep learning methods to analyze pitch, prosody, and spectral characteristics. A recurring theme is the development of lightweight, privacy-aware architectures capable of operating efficiently on mobile or edge devices, even under limited connectivity. Many of the surveyed models prioritize real-world adaptability, exhibiting strong performance in noisy conditions, across diverse dialects, and within practical communication environments such as telecom systems and voice assistant platforms.

Future research directions which include improving emotion and context sensitivity in detection algorithms, expanding the application of federated learning to strengthen the privacy and data security. The review also emphasizes the importance of reproducibility, inclusivity, ethical AI governance, and cross-institutional collaboration to promote transparency and global standardization within the deepfake detection landscape.

As a literature-based study rather than implementation of system in a direct way, this review underscores the urgent need for explainable, resource-efficient, and contextually adaptive audio deepfake detection frameworks. Such systems will be essential for securing multilingual voice security and fostering public trust in the rapidly evolving area of artificial intelligence.

References

[1] R. Singh, M. Vatsa, R. Ranjan, Multimodal deepfake detection, IEEE Transactions on Information Forensics and Security (2023).

[2] R. Ranjan, L. Ayinala, M. Vatsa, R. Singh, Multimodal zero-shot framework for deepfake hate speech detection in low-resource languages, arXiv preprint arXiv:2506.08372 (2025).
URL https://arxiv.org/abs/2506.08372

[3] O. A. Shaaban, R. Yildirim, A. A. Alguttar, Audio deepfake approaches, IEEE Access (2023).
URL https://ieeexplore.ieee.org/document/10320354

[4] M. Li, Y. Ahmadiadli, X. P. Zhang, Audio anti-spoofing detection: A survey, arXiv preprint arXiv:2404.13914 (2024).
URL https://arxiv.org/abs/2404.13914

[5] B. Zhao, Z. Kang, Y. He, et al., Generalized audio deepfake detection using frame-level latent information entropy, arXiv preprint arXiv:2504.10819 (2025).
URL https://arxiv.org/abs/2504.10819

[6] L. Nguyen-Vu, T. P. Doan, K. Hong, Detecting audio deepfakes through emotional fingerprinting, in: Lecture Notes in Computer Science, Springer, 2024.
URL https://link.springer.com/chapter/10.1007/978-981-96-7005-5_29

[7] S. Saha, M. Sahidullah, S. Das, Exploring green ai for audio deepfake detection, arXiv preprint arXiv:2403.14290 (2024).
URL https://arxiv.org/abs/2403.14290

[8] G. Tahaoglu, D. Baracchi, D. Shullani, M. Iuliani, A. Piva, Deepfake audio detection with spectral features and resnext-based architecture, Knowledge-Based Systems 323 (2025) 113726. doi:https://doi.org/10.1016/j.knosys.2025.113726.
URL https://www.sciencedirect.com/science/article/pii/S0950705125007725

[9] A. Hamza, A. R. Javed, F. Iqbal, N. Kryvinska, A. Almadhor, Z. Jalil, R. Borghol, Deepfake audio detection via mfcc features using machine learning, IEEE Access PP (2022) 1–1. doi:10.1109/ACCESS.2022.3231480.

[10] K. Sreedhar, U. Varma, Analysis of rawnet2's presence and effectiveness in audio authenticity verification, in: IEEE Conference, 2024.
URL https://ieeexplore.ieee.org/abstract/document/10911359/

[11] Z. Jin, L. Lang, B. Leng, Wave-spectrogram cross-modal aggregation for audio deepfake detection, in: IEEE ICASSP, 2025.

[12] S. Borade, N. Jain, B. Patel, V. Kumar, M. Godhrawala, S. Kolaskar, Y. Nagare, P. Shah, J. 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) (2024) 281–291.
URL https://ijisae.org/index.php/IJISAE/article/view/4972

[13] Q. Zhang, S. Wen, T. Hu, Audio deepfake detection with self-supervised xls-r and sls classifier, in: Proceedings of the 32nd ACM International Conference on Multimedia, MM '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 6765–6773. doi:10.1145/3664647.3681345.
URL https://doi.org/10.1145/3664647.3681345
[14] R. Ranjan, M. Vatsa, R. Singh, Uncovering the deceptions: Analysis on audio spoofing detection, arXiv preprint arXiv:2307.06669 (2023).
URL https://arxiv.org/abs/2307.06669
[15] C. O. Mawalim, Y. Wang, S. Okada, M. Unoki, Jmad: Multilingual audio deepfakes dataset for robust and generalizable detection, Preprint (2025).
URL https://candyolivia.github.io/assets/pdf/paper/JMADv1.pdf
[16] R. M. Purohit, A. J. Shah, D. H. Vaghera, H. A. Patil, Mladdc: Multi-lingual audio deepfake detection corpus, OpenReview (2024).
URL https://openreview.net/forum?id=ic3HvoOTeU
[17] A. Pianese, D. Cozzolino, G. Poggi, et al., Deepfake audio detection by speaker verification, in: IEEE International Conference, 2022.
URL https://ieeexplore.ieee.org/document/9975428
[18] S. Kaur, M. Buhari, N. Khandelwal, P. Tyagi, K. Sharma, Hindi audio-video-deepfake (hav-df): A hindi language-based audio-video deepfake dataset, school of Engineering & Technology, BML Munjal University, Gurugram, India (2024).
URL https://arxiv.org/abs/2411.15457
[19] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, K. A. Lee, Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2023) 2507–2522. doi:10.1109/TASLP.2023.3285283.
[20] O. Shaaban, R. Yildirim, Audio deepfake detection using deep learning, Engineering Reports 7 (03 2025). doi:10.1002/eng2.70087.
[21] J. Yi, C. Wang, J. Tao, X. Zhang, Audio deepfake detection: A survey, arXiv preprint arXiv:2308.14970 (2023).
URL https://arxiv.org/abs/2308.14970
[22] N. Chakravarty, M. Dua, Improved feature extraction for hindi language audio impersonation attack detection, Multimedia Tools and Applications (2024). doi:10.1007/s11042-023-18104-9.
[23] A. R. Ambili, R. C. Roy, Multi-tasking synthetic speech detection on indian languages, in: IEEE International Conference on Signal Processing and Communications, 2022.
URL https://ieeexplore.ieee.org/abstract/document/9744221/
[24] N. M. Müller, P. Kawa, W. H. Choong, et al., Mlaad: The multi-language audio anti-spoofing dataset, in: IEEE International Joint Conference on Biometrics, 2024.
URL https://ieeexplore.ieee.org/abstract/document/10650962/
[25] T. Tran, et al., Parallelchain lab's anti-spoofing systems for asvspoof 5, in: Proc. ASVspoof 2024, 2024.
URL https://www.isca-archive.org/asvspoof_2024/tran24_asvspoof.pdf
[26] A. Javed, K. M. Malik, A. Irtaza, et al., Voice spoofing detector: A unified anti-spoofing framework, Expert Systems with Applications (2022).
URL https://www.sciencedirect.com/science/article/pii/S0957417422002330
[27] M. Taeb, I. Kola-Adelakin, H. Chi, Forensic investigation of synthetic voice spoofing detection in social apps, in: ACM Conference, 2025.
URL https://dl.acm.org/doi/10.1145/3696673.3723086
[28] D. Salvi, P. Bestagini, S. Tubaro, Synthetic speech detection through audio folding, in: ACM Workshop, 2023.
URL https://dl.acm.org/doi/10.1145/3592572.3592844
[29] B. Zhang, H. Cui, V. Nguyen, et al., Audio deepfake detection: What has been achieved and what lies ahead, Sensors (2025).
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11991371/
[30] A. J. Lakshmi, V. Sindhuja, B. Meghana, G. S. Gupta, Detection of deepfake audio using deep learning, in: 2024 9th International Conference on Communication and Electronics Systems (ICCES), 2024, pp. 1878–1882. doi:10.1109/ICCES63552.2024.10859752.
[31] R. Anagha, A. Arya, V. H. Narayan, S. Abhishek, T. Anjali, Audio deepfake detection using deep learning, in: 2023 12th International Conference on System Modeling Advancement in Research Trends (SMART), 2023, pp. 176–181. doi:10.1109/SMART59791.2023.10428163.
[32] H. Tak, M. Todisco, X. Wang, et al., Automatic speaker verification spoofing using wav2vec 2.0, arXiv preprint arXiv:2202.12233 (2022).
URL https://arxiv.org/abs/2202.12233
[33] S. A. Momu, R. Rahman Siddiqui, S. S. Shanto, Z. Ahmed, A comprehensive approach to deepfake audio detection: Using feature fusion and deep learning, in: 2024 27th International Conference on Computer and Information Technology (ICCIT), 2024, pp. 351–356. doi:10.1109/ICCIT64611.2024.11022599.
[34] V. Velumani, P. Sekar, M. Subramanian, H. Mahaveer Chand, Deepfake detection of images, ResearchGate Preprint (2024).
URL https://www.researchgate.net/publication/380854768_Deepfake_Detection_Of_Images
[35] W. Jbara, N. A.-H. Hussein, J. Soud, Deepfake detection in video and audio clips: A comprehensive survey and analysis, Mesopotamian Journal of CyberSecurity 4 (2024) 233–250. doi:10.58496/MJCS/2024/025.
[36] A. Khan, K. M. Malik, J. Ryan, et al., Battling voice spoofing: A comparative analysis, Artificial Intelligence Review

(2023).
URL https://link.springer.com/article/10.1007/s10462-023-10539-8
[37] N. M. Müller, P. Kawa, S. Hu, et al., A new approach to voice authenticity, arXiv preprint arXiv:2402.06304 (2024).
URL https://arxiv.org/abs/2402.06304
[38] S. T. Yalla, M. G. P. Raju, D. Nagaraju, Decoding voice authenticity: Deep learning and audio features, in: IEEE Conference, 2025.
URL https://ieeexplore.ieee.org/abstract/document/10914906/
[39] A. O. Vaidya, M. Dangore, V. K. Borate, N. Raut, Y. K. Mali, A. Chaudhari, Deep fake detection for preventing audio and video frauds using advanced deep learning techniques, in: 2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2024, pp. 1–6. doi:10.1109/RAICS61201.2024.10689785.
[40] O. C. Phukan, G. S. Kashyap, A. B. Buduru, Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake, arXiv preprint arXiv:2404.00809 (2024).
URL https://arxiv.org/abs/2404.00809
[41] G. Mittal, A. Jakobsson, K. O. Marshall, C. Hegde, N. Memon, Pitch: Ai-assisted tagging of deepfake audio calls using challenge-response, arXiv preprint arXiv:2402.18085 (2024).
URL https://arxiv.org/abs/2402.18085
[42] N. V. Kulangareth, J. Kaufman, J. Oreskovic, Y. Fossat, Investigation of deepfake voice detection using speech pause patterns: Algorithm development and validation, JMIR Biomed Eng 9 (2024) e56245. doi:10.2196/56245.
URL https://biomedeng.jmir.org/2024/1/e56245
[43] Z. Almutairi, H. Elgibreen, A review of modern audio deepfake detection methods: Challenges and future directions, Algorithms 15 (5) (2022) 155. doi:10.3390/a15050155.
URL https://www.mdpi.com/1999-4893/15/5/155
[44] N. Chakravarty, M. Dua, Spectrogram-resnet41 for audio spoof attack detection with indian languages, Journal of System Assurance Engineering and Management (2024). doi:10.1007/s13198-024-02550-1.
[45] Y. Hou, H. Fu, C. Chen, Z. Li, H. Zhang, J. Zhao, Polyglotfake: A novel multilingual and multimodal deepfake dataset, in: Proc. International Conference on Artificial Intelligence, Springer, 2024.
[46] A. Cohen, D. Shyrman, A. Solonskyi, Robust prosody modeling for synthetic speech detection, SSRN (2024).
URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4892094
[47] R. M. Purohit, A. J. Shah, H. A. Patil, Ggmddc: An audio deepfake detection multilingual dataset, in: Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2024.
URL http://www.apsipa2024.org/files/papers/327.pdf
[48] P. Chiddarwar, Real-time detection of ai-generated deepfake audio: A novel approach, 2024, pp. 1–5. doi:10.1109/ICTBIG64922.2024.10911062.
[49] R. Wijethunga, D. Matheesha, A. A. Noman, K. De Silva, M. Tissera, L. Rupasinghe, Deepfake audio detection: A deep learning based solution for group conversations, in: 2020 2nd International Conference on Advancements in Computing (ICAC), Vol. 1, 2020, pp. 192–197. doi:10.1109/ICAC51239.2020.9357161.
[50] R. Ranjan, B. Dutta, M. Vatsa, Faking fluent: Unveiling the achilles' heel of multilingual deepfake detection, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.
URL https://ieeexplore.ieee.org/abstract/document/10744454/
[51] M. Chitale, A. Dhawale, M. Dubey, S. Ghane, A hybrid cnn-lstm approach for deepfake audio detection, in: 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), 2024, pp. 1–6. doi:10.1109/AIIoT58432.2024.10574576.
[52] I. Altalahin, S. AlZu'bi, A. Alqudah, A. Mughaid, Unmasking the truth: A deep learning approach to detecting deepfake audio through mfcc features, in: 2023 International Conference on Information Technology (ICIT), 2023, pp. 511–518. doi:10.1109/ICIT58056.2023.10226172.
[53] B. Sarada, T. L. Sudha, M. Domakonda, B. Vasantha, Audio deepfake detection and classification, in: 2024 Asia Pacific Conference on Innovation in Technology (APCIT), 2024, pp. 1–5. doi:10.1109/APCIT62007.2024.10673438.
[54] A. Alshehri, D. Almalki, E. Alharbi, S. Albaradei, Audio deep fake detection with sonic sleuth model, Computers 13 (10) (2024). doi:10.3390/computers13100256.
URL https://www.mdpi.com/2073-431X/13/10/256
[55] L. Pham, P. Lam, T. Nguyen, H. Nguyen, A. Schindler, Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models, in: 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2), 2024, pp. 1–5. doi:10.1109/IS262782.2024.10704095.
[56] R. Ranjan, K. Pipariya, M. Vatsa, R. Singh, Synhate: Detecting hate speech in synthetic deepfake audio in indic languages, arXiv preprint arXiv:2506.06772 (2025).
URL https://arxiv.org/abs/2506.06772
[57] S. Sarala, M. Suresh Reddy, N. Sai Kiran Reddy, V. Sai Sharan, Deepfake detection on social media, International Journal for Research Trends and Innovation (IJRTI) 9 (4) (2024) 288–291.
URL http://www.ijrti.org/papers/IJRTI2404040.pdf