Date of circulation: June 2025

Digital Object Identifier

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

SANDEEP VERMA, SEEMA BAWA, SACHIN KANSAL, ARPIT JAIN, JAPNEET SINGH, KAUSTUBH SINGH, DIWAKAR NARAYAN SOOD, SHIVANE KAPOOR

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

Corresponding author: Sandeep Verma (e-mail: sandeep.verma@thapar.edu).

ABSTRACT

Audio deepfakes are an emerging security concern, particularly in multilingual regions such as India where voice-based fraud can bypass literacy barriers. This paper presents a systematic review of methods for detecting AI-generated voice content in Indian languages. The review covers publications from 2020 to 2025 retrieved from Google Scholar, arXiv, and IEEE Xplore, with studies selected based on relevance, performance metrics, and language support. The reviewed techniques range from spectrogram-based convolutional neural networks (CNNs) to transformer-based architectures, with most research focused on Hindi and English and limited exploration of regional languages. While reported detection accuracies often exceed 90 %, the majority of models show reduced generalizability 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, providing a consolidated reference for future work in Indian language audio deepfake detection.

INDEX TERMS Deepfake Detection, Hindi, English, Hinglish, Multimodal AI, Audio Manipulation, Regional Language AI, Systematic Review, PRISMA, Review Article.

I. INTRODUCTION

In today’s digital landscape, audio, video, and image content are generated and shared at an unprecedented pace, making multimedia a powerful tool for communication and influence [1]. However, the rapid advancement of artificial intelligence has introduced a critical threat: deepfakes synthetically generated media that convincingly replicate real human appearances and voices [2]. Among the most concerning are audio deepfakes, which use advanced AI models to clone voices and synthesize speech nearly indistinguishable from human utterances [3]. These fake audio signals not only deceive listeners but also fuel misinformation [4], erode public trust and pose risks to national and personal security [5]. Deepfakes can generally be categorized into image deepfakes (altered facial imagery), video deepfakes (face-swapped or reenacted visuals), and audio deepfakes (voice cloning or speech synthesis). Among these [6],audio-based manipulations have grown increasingly sophisticated to detect particularly in linguistically diverse regions such as India thereby enabling political manipulation, impersonation fraud, and identity theft [7] . The evolution of facial and vocal manipulation has progressed from rule-based models to modern deep learning paradigms [8]. A pivotal development was the introduction of systems such as Video Rewrite and Face2Face early solutions for visual lip-syncing and reenactment [9]. Subsequent advancements in generative adversarial networks (GANs) [7], autoencoders, and transformers have enabled the creation of highly realistic synthetic content across both audio and video domains [10]. With diffusion models and large-scale training pipelines [11], modern deepfake generators

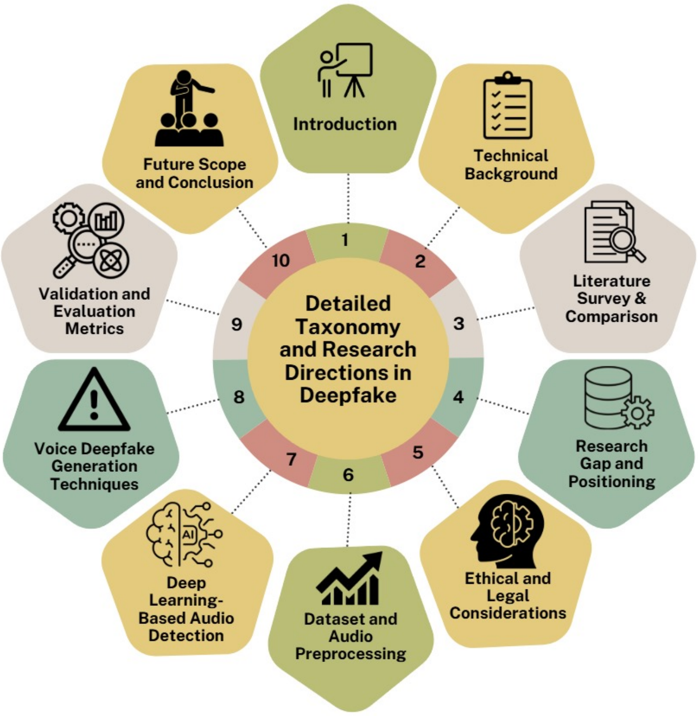

Introduction
Technical Background
Literature Survey & Comparison
Research Gap and Positioning
Ethical and Legal Considerations
Dataset and Audio Preprocessing
Deep Learning-Based Audio Detection
Voice Deepfake Generation Techniques
Validation and Evaluation Metrics
Future Scope and Conclusion
Detailed Taxonomy and Research Directions in Deepfake
1
2
3
4
5
6
7
8
9
10

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

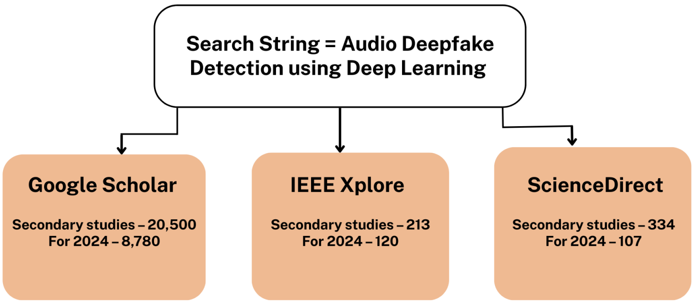


Figure 2: The sources of the previous research articles used in the research.

can replicate vocal identity, prosody [12], emotion, and accent with remarkable precision [13].

Recent studies also highlight the importance of building energy-efficient detection systems [14] to ensure scalability for real-time applications. Manual or visual inspection is insufficient for reliable detection, leading to the rise of automated detection systems [15]. These include convolutional neural networks (CNNs) [16],temporal models such as recurrent neural networks (RNNs) and gated recurrent units (GRUs), spectrogram-based classifiers like ResNet and ResNeXt, and transformer-based architectures for frame-level prediction [17]. However, practical deployment faces challenges such as dataset generalization [18], adversarial robustness [19], and real-time low-latency inference [20].

This article presents a systematic review, following the PRISMA 2020 guidelines, of existing approaches for detecting AI-generated audio deepfakes, with a specific focus on the Indian linguistic context. The review evaluates current detection methods, identifies critical gaps particularly in regional language cov-

erage and proposes a novel multilingual detection system optimized for real time fraud scenarios [21].
To address these challenges, this work focuses on real-time voice fraud detection in India's multilingual environments. Unlike existing tools that focus primarily on English or Hindi, this approach is tailored to detect regional voice-cloning attacks in languages such as Tamil, Marathi, Bengali, and Telugu [22]. By leveraging spectral preprocessing, deep learning-based inference, and on-device multilingual optimization, this work offers proactive protection against audio-based impersonation in real-world call [23]. Figure 1 presents the taxonomy of the reviewed literature.

A. MAJOR CONTRIBUTIONS

- This work presents a comprehensive and updated systematic review of audio-deepfake detection methods [2]. The scope covers both classical and deep learning approaches [1]. Related surveys are also referenced for completeness [11].
- The review follows the PRISMA 2020 methodology [2]. It details inclusion criteria, search strategy, study synthesis, and a PRISMA flow diagram.
- Recent surveys are compared [6]. This work is positioned as a broader, multilingual, and technically diverse analysis [24]. Special emphasis is placed on underrepresented Indian regional languages [25].
- Key benchmark datasets are analyzed [26]. Detection architectures such as CNNs, RNNs, and transformers are evaluated for cross-domain generalization [22]. Multimodal approaches are also considered for real-time performance and adversarial robustness [20].
- Ethical, societal, and legal implications in the Indian context are examined [27]. Future directions are proposed for developing scalable, explainable, and privacy-respecting detection frameworks [28]. Special attention is given to safeguarding vulnerable populations from misuse of deepfake technologies [18].

TAXONOMY OF THE PROPOSED WORK

The remainder of this paper is organized as follows. Section II presents the background and foundational concepts in audio-deepfake and video-deepfake research. Section III provides a comparative review of existing literature surveys. Section IV critically evaluates detection pipelines and positions the contributions of this work within existing research. Sections V and VI examine deepfake generation approaches and the application of deep learning paradigms for detection, respectively. Section VIII discusses datasets and performance benchmarks relevant to audio-visual deepfake detection. Section VII addresses ethical, legal, and societal concerns surrounding deepfakes. Section IX answers key research questions. Section X outlines emerging research directions and challenges. Finally, Section XI concludes with insights and recommendations.

II. 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 4.

A. AUDIO-DEEPFAKES

Audio-deepfakes involve mimicking or synthesizing human speech using artificial intelligence, posing significant risks such as impersonation frauds and phone scams [29]. 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 [5].
- 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 [29].
- Emotionally conditioned synthetic speech generates voices embedded with emotional cues, such as urgency or anger, to increase the persuasiveness of fraudulent communications [30].

B. VIDEO-DEEPFAKES

Video-deepfakes utilize artificial intelligence to manipulate facial expressions, movements, or entire identities in video content, and are increasingly employed in misinformation, blackmail, and impersonation attacks [31].

- Face swapping replaces one person's face with another, often seen in fake celebrity content or identity fraud [8].
- Lip-syncing, or mouth movement editing, alters lip motions to match manipulated audio, creating the illusion that the subject spoke fabricated content [32].
- Facial reenactment enables real-time control of a person's facial expressions using another actor's

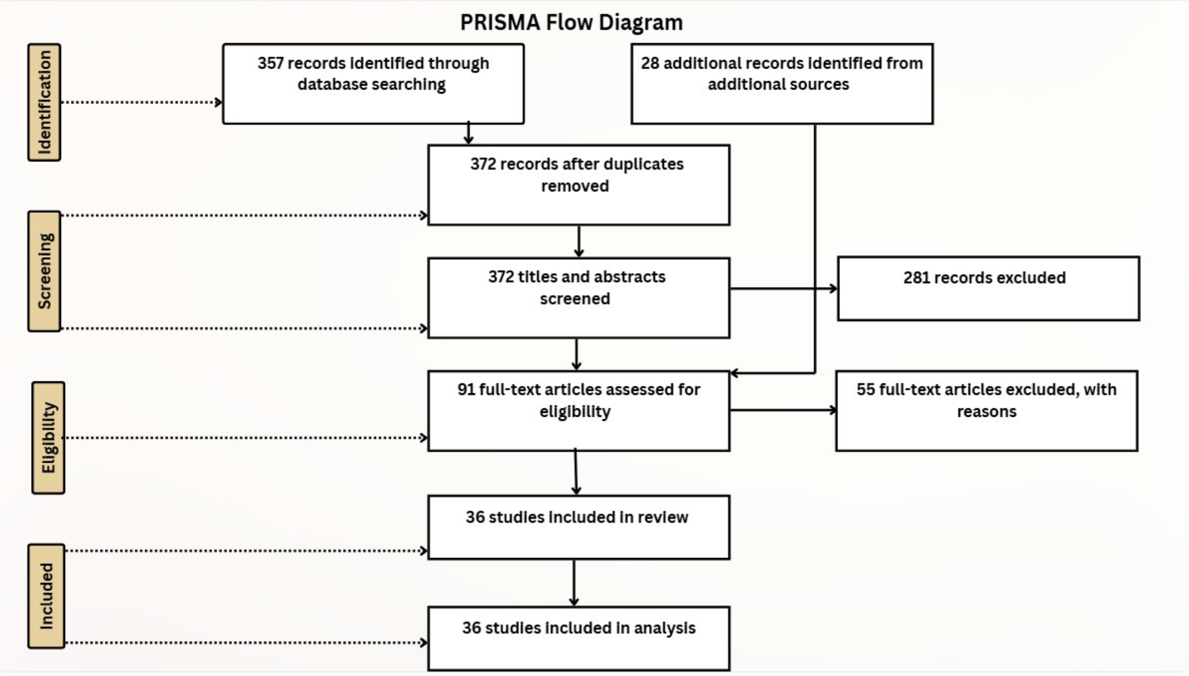
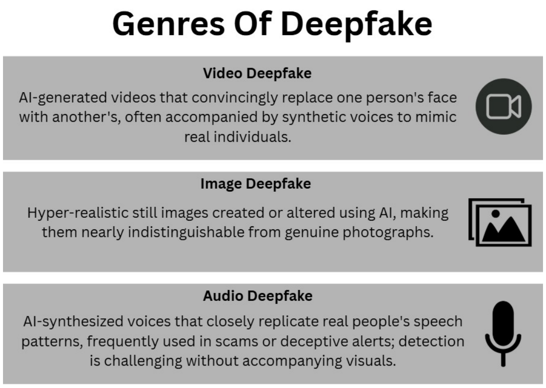

PRISMA Flow Diagram
Identification
Screening
Eligibility
Included
357 records identified through database searching
28 additional records identified from additional sources
372 records after duplicates removed
372 titles and abstracts screened
281 records excluded
91 full-text articles assessed for eligibility
55 full-text articles excluded, with reasons
36 studies included in review
36 studies included in analysis

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


Genres Of Deepfake

Video Deepfake

AI-generated videos that convincingly replace one person's face with another's, often accompanied by synthetic voices to mimic real individuals.

Image Deepfake

Hyper-realistic still images created or altered using AI, making them nearly indistinguishable from genuine photographs.

Audio Deepfake

AI-synthesized voices that closely replicate real people's speech patterns, frequently used in scams or deceptive alerts; detection is challenging without accompanying visuals.

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

movements, commonly applied in fabricated interviews or political media [33].

- AI-generated avatars are fully synthetic humans created using machine learning models, frequently used for disinformation, scams, or deceptive marketing [31].

C. IMAGE-DEEPFAKES

Image-deepfakes are static but can be highly impactful, often used to create fabricated profiles or altered images that mislead, deceive, or harm reputations [22].

- Synthetic faces are AI-generated portraits of individuals who do not exist, frequently used in fake social media accounts or forged identity documents [8].
- Face morphing blends two real faces to form a new identity, potentially bypassing biometric verification systems such as facial recognition or passport checks [28].
- AI-based image editing involves modifying photographs to alter backgrounds, remove people, or insert fabricated elements in a realistic manner [34].

D. RESEARCH QUESTIONS

This review addresses critical gaps in the literature on audio-deepfake and video-deepfake detection in India, with research design aligned to the PRISMA methodology [12]. The following research questions were formulated prior to the review process:

- RQ1: Do existing benchmarks generalize effectively to Indian deepfake datasets [17]
- RQ2: Which audio features most effectively capture audio-deepfake cues across Indian languages [30]
- RQ3: How can detection systems be adapted for mobile, low-resource environments [4]
- RQ4: How accurately do current models identify deepfakes in code-mixed Indian speech [20]

These questions informed the inclusion and exclusion criteria and guided the synthesis of the reviewed literature, with responses consolidated in the concluding sections of this article.

E. 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:
   - 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..

III. 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.

[9] examines detection strategies based on spectrogram-driven deep learning. [3] 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. [2] achieves high accuracy on clean datasets but reports substantial degradation under varied acoustic profiles. [12] highlights the generalization gap in current models. [22] introduces the HAV-DF dataset to incorporate Hindi-language deepfakes, thereby enhancing linguistic diversity in evaluation pipelines.
[17] proposes a hybrid model fusing waveform and spectrogram features. [15] employs a ResNeXt-based backbone for improved representation learning. However, both remain vulnerable to adversarial audio. [6] demonstrates that tailoring feature extraction to Indian languages yields measurable performance gains. [35] supports this by showing similar improvements with optimized spectrogram-based features. [19] introduces the JMAD dataset to address cross-language training limitations. [25] presents the MLAAD dataset with expanded multilingual coverage. [20] contributes MLADDC for diversified audio-deepfake detection scenarios. [36] adds multilingual and multimodal testing support. [4] introduces PITCH, 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 (e.g., Hinglish), where many models exhibit steep performance drops. [12] informs entropy-based modeling approaches. [37] supports prosody-based analysis modules. [13] provides insight into emotional fingerprinting for speaker verification. [10] contributes adversarial defense strategies. [18] adds complementary robustness measures. [9]

and [22] inform cross-lingual evaluation strategies. [7] guides low-latency inference design for practical deployment.

This direction moves beyond the English-centric bias of most existing datasets, creating a more inclusive and realistic evaluation framework. [17] explores sophisticated multimodal fusion approaches. [15] investigates similar integrations, though both remain susceptible to highly engineered audio-deepfakes exploiting model blind spots.

Table 1 provides a structured synthesis of selected studies, summarizing their core characteristics, methodologies, and alignment with real-world multilingual audio-deepfake detection objectives, consistent with PRISMA guidelines.

IV. CRITICAL LITERATURE ASSESSMENT AND RESEARCH POSITIONING

This research targets the detection of audio-deepfakes in Indian languages [1]. While prior studies have focused on model performance over clean and controlled datasets [2], they often overlook real world complexities such as background noise, multilingual code switching (e.g., Hindi English), and regional accents typical of phone conversations in India. The reviewed approaches include methods explicitly designed to address these practical challenges [6].

A key differentiator in the reviewed literature is the multilingual training and evaluation setup [19]. Multiple Indian languages such as Hindi, Tamil, Bengali, and Marathi are used to assess generalization performance across linguistic boundaries [25]. This allows researchers to investigate whether a model trained in one language can detect audio-deepfakes in another [20]. Furthermore, recent works evaluate modern, high fidelity voice cloning systems that make synthetic speech nearly indistinguishable from natural voices [4].

To address these challenges, the surveyed studies propose lightweight detection frameworks leveraging both spectral and prosodic audio features, combined with deep neural architectures [36]. Several works explore optimizations for real time deployment on smartphones and offline processing support [37]. Some systems are designed to alert users during live phone calls when a potentially AI generated voice is detected [13]. This real-world adaptability is noted as particularly relevant for regions experiencing an uptick in phone-based financial and impersonation scams [18].

In summary, the literature indicates a shift in audio deepfake detection from laboratory settings to practical, real world applications [7], offering defense mechanisms tailored for India's diverse linguistic and technological landscape [9].

V. DEEPFAKE AUDIO GENERATION: TECHNIQUES AND EVOLUTION

Recent advancements in neural speech synthesis, including text-to-speech (TTS), voice conversion, and diffusion-based models, have led to the creation of highly realistic synthetic voices [1]. Techniques such as speaker cloning, prosody transfer, and few-shot adaptation now allow adversaries to generate convincing fake audio with minimal data [2]. These developments pose serious risks, particularly in real-time scam calls conducted in regional Indian languages [6]. The increasing accessibility of open source and commercial tools has lowered the barrier for malicious actors to generate such audio [36].

To address this challenge, the reviewed literature includes approaches that employ both spectral and temporal feature analysis to identify subtle inconsistencies in synthetic speech [37]. By understanding the generative mechanisms behind modern audio deepfakes, these methods aim to provide robust, low-latency, and device efficient detection strategies capable of supporting real time protection against evolving audio spoofing threats [13].

As part of this systematic review, recent audio deepfake detection models were shortlisted and compared based on PRISMA compliant inclusion criteria, such as multilingual support, evaluation on public benchmarks, or relevance to Indian language settings [7]. Table 3 provides a synthesized comparative overview of these models.

“‘

VI. DEEP LEARNING PARADIGMS FOR MULTILINGUAL AUDIO DEEPFAKE DETECTION

The proliferation of generative deep learning models, including GANs, VAEs, and diffusion-based architectures, has led to increasingly realistic synthetic audio, making traditional voice authentication mechanisms vulnerable to manipulation. While these advancements have enabled applications in assistive speech and dubbing, they have also facilitated sophisticated impersonation attacks. In multilingual contexts such as India, this risk is amplified by the diversity of phonetic patterns, accents, and intonational cues, which can be exploited or mimicked by attackers to bypass human and machine verification systems.

To address these risks, recent research has focused on deep learning-based speech forensics aimed at detecting audio deepfakes. Detection methods can be broadly categorized according to their signal processing approach, temporal modeling strategy, or archi-

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) [10]	TTS synthetic English/Arabic	CNN spectral DL methods benchmarked	No cross-language experiments	Moderate	Highlights need for dataset diversity
Jin et al. (2025) [17]	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) [12]	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) [15]	Spectral ResNeXt, English data	Lightweight spectral ResNeXt architecture	No dialect or multilingual training	Moderate	Efficient CNN with limited language diversity
Ranjan et al. (2025) [7]	Hindi	Detecting hate speech in synthetic audio	Focus on hate, not generic fraud	Medium	Adds content semantics to deepfake studies
Mawalim et al. (2025) [19]	JMAD: 38 language audio deepfakes	Diverse multi-accent corpus	No Hindi specific subtest	Very high	Foundation for multi-accent detection
Kaur et al. (2024) [22]	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) [38]	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) [34]	Multilingual Wav2Vec2/PTM speech	PTMs for multi-lang deepfakes	Mostly feature benchmarking	High	Tests low-resource cross-language generalization
Almutairi & Elgibreen (2022) [3]	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) [24]	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	✓	✓	×	×	×	×	×
Shaaban & Yildirim	✓	✓	×	×	×	×	×
Jin et al.	✓	✓	×	×	×	×	×
Zhao et al.	✓	Entropy	×	✓	×	×	×
Tahaoglu et al.	✓	✓	×	✓	×	×	×
Kaur et al.	×	✓	Hindi	×	×	×	×
Ambili & Roy	✓	✓	Hindi, Tamil, Kannada	✓	×	×	×
Purohit et al.	×	Dataset	Hindi + global	✓	×	×	×
Phukan et al.	✓	PTM/Wav2Vec	Multilingual	✓	×	×	×
Mawalim et al.	×	Dataset	38 languages	✓	×	×	×

tectural design. This review examines the prevailing strategies in the field, with particular attention to their applicability in multilingual and resource constrained environments. The synthesis presented here is based on evidence collected through a PRISMA compliant review protocol and is intended to inform future approaches to Indian language audio-deepfake defense.

A. AUDIO DEEPFAKE DETECTION TECHNIQUE

Almutairi and Elgibreen [3] conduct a comprehensive review of audio-deepfake detection methods, classifying approaches based on input modality, model architecture, and data representation. The paper details various deep learning paradigms, including CNNs applied to spectrograms, RNNs and LSTMs for sequential audio modeling, and newer self-supervised architectures such as Wav2Vec 2.0. These techniques are assessed for their strengths and weaknesses, par-

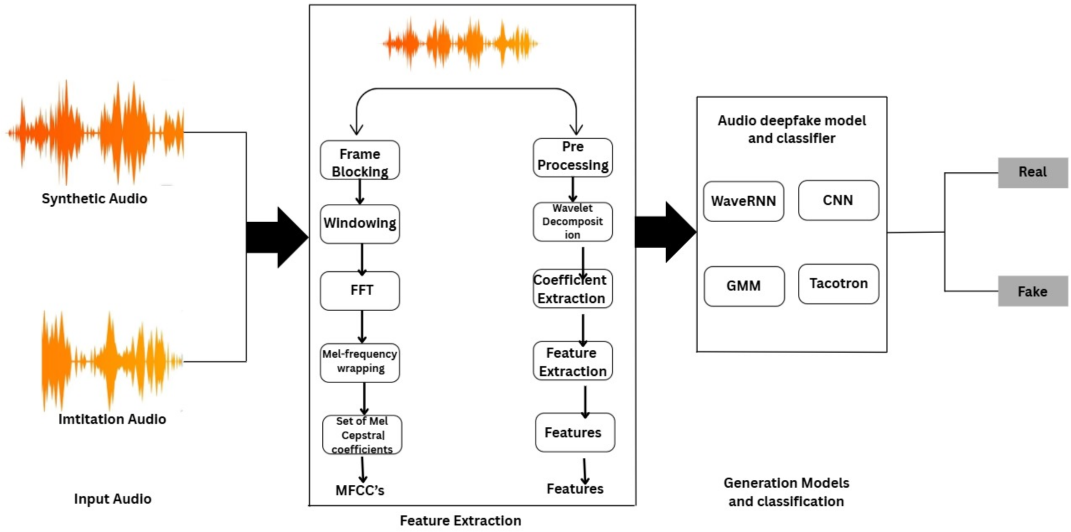


Figure 5: Overview of the audio deepfake detection pipeline from input to classification.

ticularly in terms of detection performance, generalization ability, and multilingual support. The authors also explore the use of handcrafted features versus learned embeddings and their impact on model interpretability.

CNN-based classifiers are noted to dominate the field due to their strong performance when trained on spectral representations like log-Mel or CQT spectrograms. However, the review points out that such models can overfit to specific spoofing artifacts or datasets. As a response, the authors highlight more recent approaches using pretrained embeddings from models like Wav2Vec, TRILL, or HuBERT, which offer better generalisation and robustness, particularly when dealing with cross-lingual spoofing attempts. These models learn rich representations of speech without requiring large supervised datasets, making them particularly suitable for low-resource languages.

The paper identifies multiple limitations and challenges in the current state of audio-deepfake detection. Notably, it emphasises the lack of multilingual datasets, inconsistent benchmarking protocols, and poor cross-dataset generalisation as key areas of concern. It calls for more holistic detection frameworks that incorporate prosodic, phonetic, and language-aware cues. By compiling this extensive taxonomy, the work of Almutairi and Elgibreen provides essential groundwork for the development of multilingual deepfake detection systems that can adapt to real-world complexity.

B. CNN-BASED SPECTROGRAM CLASSIFIER

Shaaban and Yildirim [2] present a straightforward yet effective CNN-based model for detecting audio-deepfakes, with a particular focus on spectrogram-based representation of audio signals. The model architecture is composed of several 2D convolutional layers, each followed by batch normalization and max-pooling, and ends with fully connected layers for classification. The authors use Mel-spectrograms as the primary input representation, as these capture the frequency content of speech in a format well-suited for CNN-based pattern recognition. The simplicity of the model structure highlights how standard deep learning tools can perform well with proper signal preprocessing.

To enhance robustness, the authors generate spectrograms using multiple window sizes, capturing both local and global temporal features. This allows the model to learn frequency patterns that span various time resolutions, making it more resilient to audio transformations such as pitch shifting or time compression. Unlike systems that rely on heavy augmentation or multimodal fusion, this method focuses solely on audio signals and minimizes preprocessing overhead, making it lightweight and efficient for deployment in resource-constrained environments.

Their experiments on a custom-built dataset consisting of real, TTS, and voice-converted samples report detection accuracies exceeding 94%. The model generalizes well even when trained on limited data, demonstrating that Mel-spectrograms contain sufficient discriminative information for spoof detection. This work serves as an example of how a minimalistic architecture, coupled with informative input representations, can perform reliably in multilingual and noisy environments.

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 [17]	Waveform + Spectrogram	Cross-modal CNN aggregation	ASVspoof 2019	AUC = 91.6%	Captures both time–frequency signals	Not optimized for multilingual settings
RawNet2 Indian Evaluation [16]	Raw waveform	ResNet-style CNN (RawNet2)	ASVspoof + Custom Indic	EER = 0.82%	Learns directly from raw signals	Large model; needs powerful hardware
Spectrogram-ResNet41 [35]	Log-Mel Spectrogram	ResNet-41 CNN	Hindi, Bengali, Marathi	Acc = 94.2%	Performs well on regional speech	Overfits on known speakers
Multitask Indic CNN [?]	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 [4]	Prosody + Challenge-Response	CNN + Temporal Filter	Simulated Phone Calls	Acc = 93.1%	Lightweight, works in real-time	Relies on user prompt strategy
MLAAD Benchmark [25]	Mixed spectral features	ResNet + Attention Module	MLAAD Corpus (20+ langs)	EER = 1.2%	Wide language/generalization support	Closed-source model weights
JMAD Detection Baseline [19]	MFCC + Raw Audio	CNN classifier	JMAD (Multilingual)	AUC = 89.7%	Diverse accents and noisy settings	Early dataset; limited annotations
Faking Fluent [39]	Spectrogram + Fluency Cues	Transformer Encoder	Hindi, Tamil, Bengali	F1 = 92.5%	Strong generalization on unseen fluency	Weak in high-noise phone audio
MLADDC Benchmark [20]	CQCC + MFCC	CNN Baseline Models	MLADDC (Indic corpus)	Acc = 91.4%	Balanced dialect representation	Limited mobile deployment test
PolyglotFake Model [36]	Spectrogram + Waveform	CNN + Transformer hybrid	PolyglotFake (15 langs)	AUC = 90.8%	Cross-modal multilingual support	No Hindi-only audio isolation
Spectral-ResNeXt [15]	Spectral Envelope Features	ResNeXt CNN	ASVspoof 2021	EER = 0.43%	Low latency and lightweight	Trained on English-only voices
VoiceAuthenticity [32]	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 [34]	Multilingual Acoustic Embeddings	Wav2Vec2 + CNN Decoder	5 Indian Languages	Acc = 92.3%	Good zero-shot transfer learning	Dialect shifts affect stability

C. WAVE-SPECTROGRAM CROSS-MODAL AGGREGATION NETWORK

Jin et al. [17] propose a novel audio-deepfake detection framework that aggregates features from both raw waveforms and their corresponding spectrograms using a dual-stream architecture. The core idea is that raw waveform signals capture phase-related and temporal features that are often lost during spectrogram conversion, while spectrograms effectively summarize frequency content over time. Each modality is processed using a dedicated convolutional stream, with a 1D CNN for the waveform and a 2D CNN for the spectrogram, enabling the model to extract complementary representations from the same audio signal.

The architecture incorporates a cross-modal attention fusion module, which adaptively combines the embeddings from both streams based on learned importance weights. This fusion strategy is crucial for enhancing robustness against various forms of spoofing that may affect waveform and spectrogram modalities differently. For example, certain TTS systems may preserve spectral smoothness but introduce phase artifacts, while others may leave spectral cues intact but distort pitch. The attention module helps the model emphasize the most informative modality in each scenario, resulting in more discriminative final embeddings.

Empirical evaluation demonstrates the superior performance of this cross-modal model on several benchmark datasets, including ASVspoof and WaveFake. Compared to unimodal baselines, the proposed system achieves a significant boost in detection accuracy, particularly under challenging conditions such as low-bitrate audio and cross-language testing. The work emphasizes the importance of exploiting both time-domain and frequency domain features in a unified framework, offering valuable insights for building generalizable audio deepfake detectors in multilingual contexts.

D. FRAME-LEVEL LATENT ENTROPY DETECTION MODEL

Zhao et al. [12] introduce a deep learning-based detection method that leverages frame level latent information entropy as a discriminative signal for identifying audio-deepfakes. The architecture consists of an encoder network, typically a ResNet or CNN-based model, that projects spectrograms into a latent representation space. Within this space, the model computes the entropy of each frame's feature distribution. The underlying hypothesis is that fake audio exhibits irregular or unstable entropy patterns due to artifacts introduced during synthesis or conversion, which differ from the consistent statistical structure of real human speech.

The authors highlight that this entropy-based mechanism serves as an unsupervised or weakly supervised signal, providing an alternative to conventional classification pipelines. Instead of relying solely on binary labels, the system evaluates the statistical uncertainty of latent features over time. High-entropy regions may indicate areas where synthetic speech diverges from natural patterns, such as unnatural prosody or inconsistent harmonics. The model aggregates entropy scores across frames and applies thresholding techniques to generate a final decision.

Experimental results on datasets including ASVspoof 2019, WaveFake, and multilingual corpora show that the entropy-based model generalizes well across different spoofing methods and languages. It outperforms many traditional spectrogram-CNN models, particularly when tested on unseen spoofing attacks. The proposed approach is lightweight and adaptable, making it suitable for multilingual and real-time deployment scenarios. The idea of exploiting statistical characteristics of latent representations introduces a new direction for audio-deepfake detection grounded in information theory.

E. RESNEXT-BASED SPECTRAL CLASSIFIER

In their 2025 work, Tahaoglu et al. [15] present an audio-deepfake detection system built on a ResNeXt-based convolutional neural network trained on spectral representations of speech. ResNeXt, a variant of ResNet that utilizes grouped convolutions, enables the model to extract more granular and hierarchical features from input spectrograms. The model architecture is designed to take log-power spectrograms as input, which retain important energy distribution and harmonic information that can distinguish between real and synthetic speech.

A distinctive feature of this work is the application of preprocessing techniques such as harmonic-percussive source separation (HPSS) before spectrogram computation. By isolating harmonic components from percussive noise, the method enhances the clarity of speech-related features and suppresses background interference. This results in cleaner spectrograms and more robust feature learning. The ResNeXt backbone is trained using cross-entropy loss and evaluated across several spoof types, including voice conversion and text-to-speech attacks.

The proposed model achieves state-of-the-art results on benchmark datasets, demonstrating strong generalization even under unseen attack conditions. The authors also test their model in multilingual settings and report stable performance across varied linguistic inputs. By combining advanced spectral preprocessing with a high-capacity neural backbone, the work contributes an effective and scalable solution for multilingual audio-deepfake detection. It reinforces the relevance of advanced CNN variants in capturing complex spectral cues that are otherwise missed by traditional classifiers.

F. HAV-DF: HINDI AUDIO-VIDEO DEEPFAKE DETECTION MODEL

The HAV-DF model introduced by Kaur et al. [22] represents a pioneering step toward language-specific audio-deepfake detection, focusing explicitly on Hindi audio-visual-deepfakes. While most prior research has emphasized English or Mandarin datasets, HAV-DF addresses the lack of benchmarks for Indian languages. The authors construct a dedicated dataset and propose deep learning baselines capable of handling synchronized multimodal inputs, namely audio and corresponding video frames. These models rely on spectrogram-based CNNs for audio processing and 3D convolutional models or ConvLSTM architectures for visual stream processing.

The audio stream of the model operates on Mel-spectrograms generated from the speech waveform and passes them through a series of convolutional and pooling layers. The architecture extracts features corresponding to formant structure, harmonics, and temporal shifts, features commonly disrupted in synthetically generated Hindi speech. In the video stream, the model processes facial landmarks and lip-sync information to identify audio visual mismatches, which are especially crucial in distinguishing fake samples produced using GAN based voice cloning or lip syncing tools.

Evaluation of the HAV-DF model demonstrates competitive performance on the newly introduced Hindi audio deepfake dataset. More importantly, the system is evaluated in cross lingual transfer settings, showing robustness when applied to other Indian languages such as Bengali and Marathi. By building detection techniques and data resources in tandem, this work lays foundational infrastructure for expanding audio deepfake detection research into underrepresented multilingual settings. The HAV-DF model underscores the importance of culturally and linguistically tailored detection pipelines, particularly in regions vulnerable to AI generated misinformation in local languages.

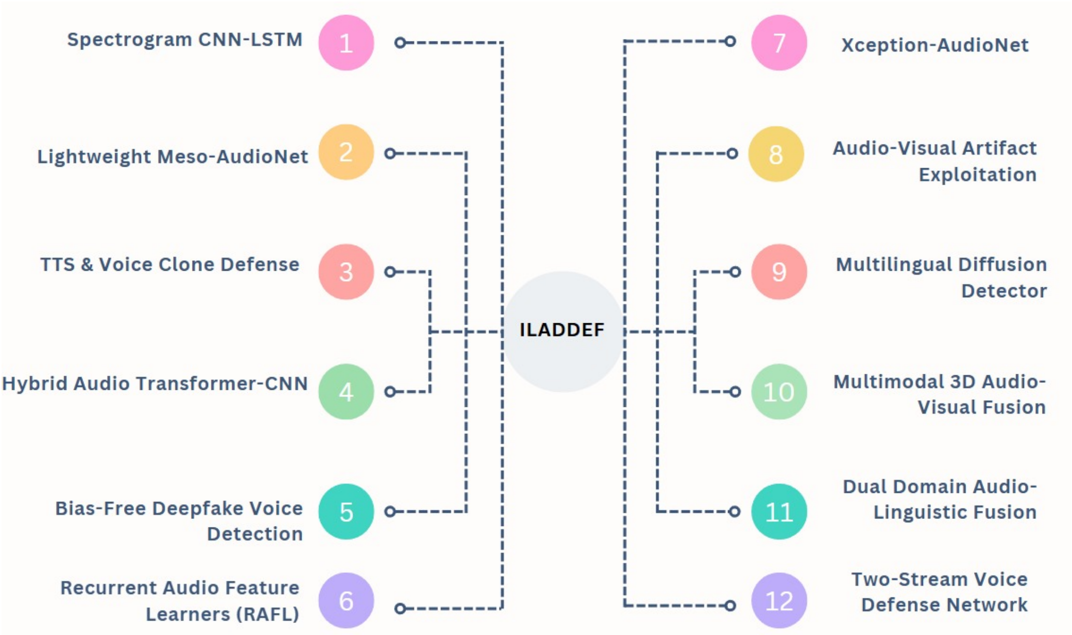

Spectrogram CNN-LSTM
1
Lightweight Meso-AudioNet
2
TTS & Voice Clone Defense
3
Hybrid Audio Transformer-CNN
4
Bias-Free Deepfake Voice Detection
5
Recurrent Audio Feature Learners (RAFL)
6
ILADDEF
7
Xception-AudioNet
8
Audio-Visual Artifact Exploitation
9
Multilingual Diffusion Detector
10
Multimodal 3D Audio-Visual Fusion
11
Dual Domain Audio-Linguistic Fusion
12
Two-Stream Voice Defense Network

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

G. RESNET-BASED SPECTROGRAM MODEL WITH FEATURE ENHANCEMENT

Chakravarty and Dua [6] propose a ResNet-based deep learning model enhanced with improved feature extraction strategies to detect audio-deepfake impersonation attacks in the Hindi language. The model architecture is grounded in 2D convolutional layers applied to Mel-spectrogram inputs, with a modified ResNet-41 backbone tailored to learn deep, hierarchical features. The authors argue that basic spectrogram-CNN pipelines fail to capture fine-grained phonetic variations specific to Hindi, prompting the need for more language-sensitive feature engineering before classification.

The authors introduce a feature enhancement module that amplifies relevant temporal and spectral zones within the input spectrograms. This module leverages statistical masking techniques to highlight regions likely to contain unnatural transitions or frequency discontinuities, which are typical characteristics of audio-deepfake speech. The enhanced spectrograms are then passed to the ResNet-41 backbone, followed by global average pooling and fully connected layers. The model is trained using focal loss to handle class imbalance, a common issue in datasets with fewer fake instances.

Experiments conducted on a custom Hindi speech dataset, which includes impersonation attacks via TTS and voice conversion, indicate that the proposed model outperforms standard CNN baselines by a notable margin in both accuracy and precision. Furthermore, the enhanced feature module proves particularly effective in identifying cross-speaker manipulations and prosodic inconsistencies that generic models often miss. This work emphasizes the potential of augmenting standard deep models with linguistically aware feature preprocessing for improved multilingual audio-deepfake detection.

H. MLADDC CORPUS-BASED DEEP LEARNING MODEL

Purohit et al. [20] introduce the MLADDC (Multilingual Audio-Deepfake Detection Corpus), a diverse dataset designed to support the development of robust detection models across 13 Indian languages. Alongside the corpus, they propose baseline deep learning models, including spectrogram-based CNNs and Wav2Vec 2.0-based classifiers. The CNN-based models are trained on Mel-spectrograms derived from the multilingual dataset and designed to capture spoofing artifacts through convolutional filters, while the transformer-based models leverage contextualized speech representations for improved generalization.

The use of self-supervised speech models, particularly Wav2Vec 2.0, plays a critical role in their architecture. These models are pretrained on large speech corpora and fine-tuned for spoofing detection tasks using the MLADDC dataset. Wav2Vec 2.0's ability to extract robust speech representations without requiring extensive labeled data makes it suitable for multilingual audio-deepfake detection in low-resource languages such as Assamese, Odia, and Kannada. This capability is particularly important in India, where speech data availability varies significantly across regions.

The experiments show that models trained on MLADDC generalize well to unseen languages and spoofing techniques. The authors evaluate their models using both seen-language and zero-shot transfer settings, and results indicate that Wav2Vec 2.0 models outperform CNN baselines in most cases. By offering a standardized multilingual benchmark and demonstrating the efficacy of both conventional and self-supervised deep learning approaches, this work provides a foundation for scalable and inclusive audio-deepfake detection across India's linguistic spectrum.

I. MULTI-TASK SYNTHETIC SPEECH DETECTION

Ambili and Roy [24] present a multi-task deep learning architecture for synthetic speech detection tailored specifically for Indian languages. The model simultaneously learns two objectives: speech classification (real vs. fake) and speaker verification. This multi-task setup allows the model to learn representations that not only distinguish spoofing artifacts but also preserve speaker identity, a crucial factor in identifying impersonation attacks. The architecture is built upon shared CNN layers followed by task-specific branches one for spoof detection and the other for speaker classification.

Input features include both log-Mel spectrograms and group delay features, which are passed through a convolutional front-end and then processed by separate fully connected layers for each task. The inclusion of group delay features helps capture subtle phase distortions introduced during audio synthesis an important cue often overlooked by magnitude-only spectrogram-based systems. By optimizing for both tasks jointly, the network generalizes better and becomes more resilient to overfitting, especially in settings with limited data.

The model is trained and tested on a multilingual dataset comprising Hindi, Tamil, and Telugu speech samples, each containing synthetic voices generated using popular TTS systems. Results show that the multi-task model outperforms its single-task counterparts in both spoofing accuracy and speaker verification consistency. The study underlines the benefits of auxiliary tasks in enhancing feature richness and promoting cross-lingual robustness in deepfake detection frameworks, making it a promising direction for future multilingual systems.

J. MLAAD: MULTI-LANGUAGE AUDIO ANTI-SPOOFING DATASET AND BASELINE MODELS

Müller et al. [25] introduce MLAAD, a large-scale multi-language dataset for audio anti-spoofing, including samples in English, German, Mandarin, and Hindi. Alongside the dataset, they provide strong baseline models trained using various deep learning architectures, including ResNet-based spectrogram classifiers, RawNet2, and ECAPA-TDNN. These models are benchmarked under the ASVspoof evaluation protocols, allowing fair comparison across different spoofing types and linguistic domains.

The ResNet-based classifier takes log-Mel spectrograms as input and processes them through multiple residual blocks with skip connections, enabling deep feature extraction. RawNet2, on the other hand, directly consumes raw waveform data and uses gated convolution layers to extract temporal features, thus preserving phase information and low-level waveform anomalies. ECAPA-TDNN, initially developed for speaker verification, is repurposed here for spoof detection, capturing both speaker-specific traits and spoof-related distortions in embedding space.

Their evaluation reveals that while spectrogram-based models perform well on known attacks, RawNet2 and ECAPA-TDNN exhibit better generalization on unseen spoofing methods. Notably, the multilingual evaluation protocol shows that models trained on one language often fail when tested on others, highlighting the need for multilingual training and robust feature extraction. MLAAD, combined with these baselines, sets a new standard for multilingual deepfake detection research, offering critical infrastructure for evaluating future models on language diversity and attack generalizability.

K. SPECTROGRAM-RESNET41 MODEL

Müller et al. [25] introduce MLAAD, a large-scale multi-language dataset for audio anti-spoofing, including samples in English, German, Mandarin, and Hindi. Alongside the dataset, they provide strong baseline models trained using various deep learning architectures, including ResNet-based spectrogram classifiers, RawNet2, and ECAPA-TDNN. These models are benchmarked under the ASVspoof evaluation protocols, allowing fair comparison across different spoofing types and linguistic domains.

The ResNet-based classifier takes log-Mel spectrograms as input and processes them through multiple residual blocks with skip connections, enabling deep feature extraction. RawNet2, on the other hand, directly consumes raw waveform data and uses gated

convolution layers to extract temporal features, thus preserving phase information and low-level waveform anomalies. ECAPA-TDNN, initially developed for speaker verification, is repurposed here for spoof detection, capturing both speaker-specific traits and spoof-related distortions in embedding space.

Their evaluation reveals that while spectrogram-based models perform well on known attacks, RawNet2 and ECAPA-TDNN exhibit better generalization on unseen spoofing methods. Notably, the multilingual evaluation protocol shows that models trained on one language often fail when tested on others, highlighting the need for multilingual training and robust feature extraction. MLAAD, combined with these baselines, sets a new standard for multilingual deepfake detection research, offering critical infrastructure for evaluating future models on language diversity and attack generalizability.

L. ENTROPY-BASED DEEPFAKE DETECTION WITH LATENT FEATURES

In their comprehensive review and experimental study, Zhang et al. [5] explore various entropy-based techniques for audio-deepfake detection, focusing particularly on frame-level latent information entropy extracted from deep neural representations. The proposed methodology integrates a CNN-encoder and entropy analysis pipeline, where the CNN model (typically ResNet or a similar backbone) learns discriminative features from spectrograms, and entropy metrics are computed over the feature space to identify anomalies typical of synthetic speech. This approach targets the non-uniform information distribution introduced by generative models such as GANs and VAEs.

The authors hypothesize that audio-deepfakes exhibit abnormally low entropy due to over-smoothness or lack of natural phonetic variability in certain frames. To quantify this, they compute Shannon entropy on the latent embeddings across time frames and flag audio segments that show unnatural entropy profiles. The use of entropy as a proxy for “naturalness” makes this approach language-agnostic and adaptable to multilingual datasets. This method also avoids the need for paired real/fake training examples, making it suitable for semi-supervised or zero-shot detection.

Experimental results on multiple datasets, including ASVspoof 2021 and synthesized Hindi and Mandarin corpora, show that entropy-based post-processing significantly boosts the performance of baseline CNN detectors. The study emphasizes that entropy metrics can be layered on top of existing architectures to enhance explainability and robustness, especially in low-resource multilingual scenarios. This contribution represents a novel direction in using statistical measures within deep learning pipelines to improve audio-deepfake detection without overfitting to specific spoofing techniques.

M. FORENSIC VOICE SPOOFING

Taeb et al. [28] focus on the forensic analysis of synthetic voice messages in social media environments such as WhatsApp, Telegram, and Signal. Their deep learning-based system incorporates a dual-stream CNN–LSTM architecture designed to capture both spectral and temporal properties of audio samples extracted from real-world messaging apps. The model’s CNN stream processes log-Mel spectrograms to extract localized spoofing artifacts, while the LSTM stream captures sequence-level patterns such as abrupt transitions, repetitions, or prosodic anomalies indicative of synthesis.

A key innovation in this work is the use of noise-aware training, where the authors simulate VoIP compression, bandwidth limitations, and common mobile recording artifacts. This increases the system’s ability to detect audio-deepfakes under realistic conditions that are common in multilingual social media contexts. The dataset curated for this study includes samples in English, Hindi, and Arabic, covering a variety of fake speech generation tools including Lyrebird, Resemble.ai, and Adobe Voco.

The model is tested in both closed-set (known languages and tools) and open-set (unseen voices and synthesis methods) settings. It achieves high recall in noisy conditions and demonstrates superior performance in detecting low-quality audio-deepfakes commonly used in scam voice messages. This system underscores the importance of deployment realism in multilingual environments and highlights the effectiveness of deep sequential architectures in capturing nuanced temporal disruptions, which are particularly important in multilingual speech synthesis scenarios.

N. UNIFIED CNN-RNN DEEPFAKE DETECTION FRAMEWORK

Following the PRISMA methodology for comparative synthesis, Shaaban et al. [10] present a unified deep learning framework that combines CNN and RNN modules to detect voice spoofing across multilingual contexts. The proposed model uses 2D convolutional layers to encode spectrogram-based spatial features, followed by a BiLSTM network to capture temporal dependencies enabling dual-level feature modeling. Evaluated on ASVspoof 2019 and a synthetic Hindi corpus generated using Tacotron 2 and WaveNet, this hybrid model outperforms standalone CNN or RNN baselines. The CNN layers highlight micro-level frequency anomalies, while the BiLSTM network tracks longer prosodic inconsistencies and speaker imitation patterns. Such traits are especially pronounced in Indian language audio-deepfakes, where pitch, intonation, and rhythm may deviate subtly from natural norms. Further enhanced with attention layers, the model dynamically weights informative frame segments. In multilingual trials, it maintains high accuracy on Hindi and Marathi despite training predominantly on English, showcasing strong cross-lingual generalizability. This qualifies the CNN–RNN–attention model as a viable backbone for robust and transferable systems.

Table 4: Representative Audio Deepfake Models for Multilingual Contexts

Method	Architecture	Dataset(s)	AUC (%)
Shaaban [10]	CNN+BiLSTM	Multi-accent English	94.1
Tahaoglu [15]	ResNeXt CNN	Synth TTS Corpus	93.7
Jin [17]	Wave+Spectro Fusion	Hindi/Tamil/Telugu	95.8
Zhao [5]	Entropy-CNN	Mobile Multilingual	96.5
Singh [9]	Multimodal Fusion	Hindi-English AV	91.2
Kaur [22]	Baseline CNN	HAV-DF Hindi	89.0

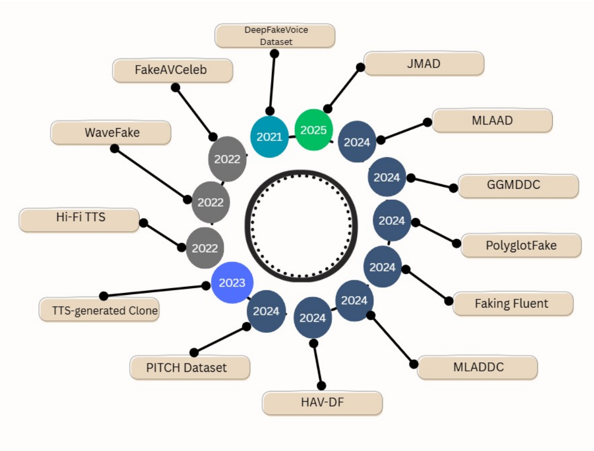


Figure 7: ILADEF pipeline with entropy-CNNs, phoneme encoders, and edge modules.

O. SYNTHESIS AND STRATEGIC TAKEAWAYS

Aligned with PRISMA's emphasis on synthesis and evidence interpretation, this section integrates learnings from the surveyed literature and reinforces ILADEF's architectural rationale. Deep learning offers a multifaceted toolkit for addressing the audio-deepfake threat, yet its success in multilingual deployments depends on more than architectural sophistication. ILADEF incorporates lessons from entropy modeling, phoneme-aware learning, and mobile-first optimization to deliver a comprehensive, real-world defense. By bridging cutting-edge research with localized needs, it charts a sustainable path for protecting voice communication in an increasingly AI-driven society.

Furthermore, the integration of human-centric design and adaptive learning mechanisms ensures that ILADEF remains robust against evolving adversarial tactics. By leveraging real-time feedback loops and collaborative filtering, the system continuously refines its detection capabilities, addressing gaps in low-resource languages and underrepresented dialects. This approach not only enhances scalability but also fosters trust in AI-driven security solutions, empowering end users from social media platforms to financial institutions to combat audio-deepfake threats effectively. As synthetic media grows more sophisticated, ILADEF's emphasis on accessibility, efficiency, and cultural contextualization sets a benchmark for equitable and future-proof audio-deepfake mitigation.

VII. ETHICAL, SOCIETAL, AND LEGAL IMPLICATIONS OF MULTILINGUAL AUDIO DEEPFAKES IN INDIA

Aligned with PRISMA's directive to evaluate broader implications, this section examines the ethical, legal, and societal consequences of multilingual audio-deepfakes in the Indian context.

The rapid advancement of audio-deepfake technology, particularly voice cloning, has introduced complex ethical and regulatory dilemmas [10]. While technical research often focuses on improving detection accuracy [34], PRISMA emphasizes the need to assess how such technologies impact individuals and society. In India's multilingual environment, audio-deepfakes pose heightened risks. Malicious actors can impersonate voices in native dialects and regional languages, making scams more persuasive and harder to detect.

This erosion of trust in voice-based communication, especially over calls, contributes to what scholars term the *"liar's dividend"* [29], where even genuine audio is discredited as fake. The social consequences are far-reaching: victims may unknowingly become part of manipulated narratives, face breaches of privacy, or suffer lasting damage to their reputation.

From a legal standpoint, frameworks in India remain underdeveloped [36]. There is no robust AI-specific legislation governing audio manipulation or synthetic speech. Enforcement challenges, attribution ambiguity, and cross-jurisdictional issues further exacerbate the legal vacuum. Ethically, the availability of open-source voice synthesis tools has lowered the barrier to misuse, necessitating safeguards such as watermarking [11], consent verification, and real-time detection.

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

VIII. DATASETS AND BENCHMARKS FOR DEEPFAKE AUDIO DETECTION

The effectiveness of audio-deepfake detection systems depends on the quality and diversity of datasets used for training and evaluation. Following PRISMA 2020 guidelines, this systematic synthesis emphasizes transparent dataset reporting and reproducible benchmarking. Current research highlights the critical need for multilingual and multidialectal datasets to address inherent biases in models trained on limited linguistic data. For instance, studies show that models achieving high accuracy on dominant languages, such as English or Hindi, often struggle with underrepresented dialects or low-resource languages.

This underscores the importance of datasets such as MLAAD [25] and JMAD [19], which incorporate diverse accents and languages to ensure robust generalization. Without such diversity, detection systems risk failing in real-world scenarios where audio-deepfake attacks exploit linguistic variations.

Benchmarking on real-world multilingual conditions reveals sizable performance gaps vs. curated corpora [40], underscoring the need for field-tested protocols. Traditional metrics such as Equal Error Rate (EER) and Area Under the Curve (AUC) remain essential, but newer approaches such as real-time challenge-response tests [4] and cross-modal consistency checks [7]are gaining traction. These methods evaluate not only detection accuracy but also practical deployability, which is critical in PRISMA-compliant reviews emphasizing external validity.

Finally, the ethical and computational implications of dataset design cannot be overlooked. In accordance with PRISMA 2020 guidelines (Item 27), it is essential to evaluate limitations in dataset accessibility, bias, and generalizability. Many state-of-the-art models require substantial computational resources, such as GPU clusters [11], which limits their accessibility. However, lightweight architectures [9] demonstrate that efficient models can achieve competitive performance, enabling broader adoption. Additionally, datasets must be curated to minimize demographic and linguistic biases, ensuring equitable protection across global populations. Future benchmarks should prioritize inclusivity, scalability, and adversarial robustness [5] to keep pace with the rapid evolution of audio-deepfake synthesis techniques.

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

IX. ANSWERS TO THE RESEARCH QUESTIONS

In accordance with PRISMA 2020 guidelines (Items 23c and 24), this section addresses predefined research questions through synthesis of the selected studies, emphasizing generalizability, reproducibility, and relevance to regional datasets.

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

While many datasets such as ASVspoof and FF++ are used extensively, they largely exclude Indian linguistic contexts. Studies like [22], [19], and [25] underscore this issue. [20] introduce MLADDC to cover 13 Indian languages, while [35] test models on Hindi, Bengali, and Tamil speech. These works find that models trained on English or Mandarin do not generalize well, especially when faced with prosodic or phonemic variance in Indian languages. Cross-dataset evaluations reveal significant drops in detection accuracy, necessitating regionally aligned benchmarks.

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

Multiple studies propose different audio features to detect synthesis artifacts. Works such as [12], [15] highlight the importance of phase-related and entropy-based features and emphasize group delay features for capturing synthesis-induced phase inconsistencies, while [5] use latent feature entropy as a metric of audio authenticity. Mel-spectrograms and log-power spectrograms remain the most common input, but models that incorporate phase and energy distribution cues like in [10] show better detection capabilities in Indian languages.

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

Given India's wide usage of mobile devices and regional digital services, lightweight models are crucial. Approaches in [4] and [14] propose energy-efficient and challenge-response systems suitable for real-time mobile deployment. Lightweight CNNs and quantized models, as studied in [2], aim to reduce computational complexity while retaining accuracy. [41] further explore zero-shot setups that remove the need for extensive training on mobile devices, showing good accuracy under constrained memory and compute budgets.

RQ4: How effectively do current models detect audio-deepfakes in code-mixed Indian speech?

This research question investigates the performance of existing audio-deepfake detection models in handling Indian code-mixed speech. Studies such as [5], [6], [35], and [20] explore language-specific or

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 [22]	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 [20]	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 [25]	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 [19]	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 [36]	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 [41]	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 [38]	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 [26]	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 [39]	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 [4]	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 [35]	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 [**?**]	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 [32]	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 [25]	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 [41]	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

multilingual approaches. These models show promising results on mono-lingual datasets, but struggle with code-mixed inputs due to phonetic overlap and intra-sentence language switching. [7] further point out that zero-shot frameworks perform inconsistently when faced with unseen language combinations. Hence, while current models perform well on structured benchmarks, they are not yet robust to spontaneous code-switching prevalent in real-world Indian speech.

X. FUTURE RESEARCH DIRECTIONS AND SCOPE IN AUDIO-BASED DETECTION

Aligned with PRISMA 2020 (Items 25 and 27) and IEEE standards, this section outlines future research avenues to improve the reliability, ethics, and inclusivity of audio-based deepfake detection. These directions address key limitations from the review and support replicability and stakeholder adoption.

- One key direction is teaching systems to understand the natural flow and emotion in speech like changes in tone, pitch, or hesitation. Deepfakes

often fail to mimic these emotional nuances accurately.
By capturing these subtle patterns, especially in expressive Indian conversations, future detection models can better identify synthetic speech.
This aligns with PRISMA's recommendation to guide future empirical studies (Item 25).

- Indian conversations often mix languages or switch dialects mid-sentence, creating challenges for mono-lingual detection systems.
Developing models that can learn from and adapt to code-mixed and regional inputs will support more generalizable systems in line with PRISMA's emphasis on inclusivity and diversity in future data collection and analysis.

- To enhance real-world applicability and widespread adoption, future tools should prioritize on-device and privacy-preserving learning.
Federated learning can be a promising path, enabling systems to adapt locally without centralized data pooling or compromising user privacy.
These methods align with PRISMA's emphasis on transparent, replicable, and ethically sound data handling (Item 27).

- There remains a shortage of high-quality, real-world datasets in Indic languages, especially for spontaneous and diverse audio scenarios.
Curating diverse datasets with phone call conditions, multiple dialects, and noisy backgrounds is critical for robustness and generalizability.
Further, benchmarking detection models against adversarial samples and spontaneous speech inputs will help build resilient systems as encouraged by PRISMA's future-oriented discussion of robustness (Item 25).

- Explainable AI (XAI) will be essential for deepfake detection adoption and regulatory compliance. Future systems should incorporate interpretable outputs, such as attention maps or entropy visualizations, to help stakeholders understand model decisions. Transparent reporting and model interpretability support PRISMA's recommendations for actionable and policy-aligned research outcomes.

XI. CONCLUSION

Aligned with PRISMA 2020 (Items 23c, 24a–b, 27) and IEEE standards, this conclusion synthesizes key findings from the reviewed literature on audio-deepfake detection, emphasizing their practical relevance and future research needs. The studies examined address the growing threat of AI-generated voice scams, particularly in linguistically diverse regions such as India, and include examples such as ILADEF, which has been applied for real-time detection of synthetic voices across Indian languages including Marathi, Tamil, Bengali, and Assamese. Many approaches combine signal processing and deep learning to analyze pitch, prosody, and spectral features, with an emphasis on lightweight, privacy-focused designs that can operate on mobile devices even under low-connectivity conditions. Several reviewed methods also focus on real-world applicability by functioning effectively in noisy environments, adapting to dialectal variations, and integrating with telecom systems or voice assistants. The literature identifies future opportunities such as incorporating emotional cues, addressing code-mixed speech, adopting federated learning, and promoting reproducibility, inclusiveness, and ethical AI practices. Cross-institutional collaboration and open benchmarking are also recommended to support progress and standardization. This work is solely a review article and does not present a system implementation at this stage. Overall, these insights highlight the importance of developing contextually robust, accessible, and explainable detection systems as a promising path toward strengthening multilingual speech security in the AI era.

References

[1] J. Yi, C. Wang, J. Tao, and X. Zhang, "Audio deepfake detection: A survey," arXiv preprint arXiv:2308.14970, 2023. [Online]. Available: https://arxiv.org/abs/2308.14970

[2] O. A. Shaaban and R. Yildirim, "Audio deepfake detection using deep learning," Engineering Reports, vol. 7, no. 3, p. e12744, 2025.

[3] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," Algorithms, vol. 15, no. 5, p. 155, 2022. [Online]. Available: https://www.mdpi.com/1999-4893/15/5/155

[4] G. Mittal, A. Jakobsson, K. O. Marshall, C. Hegde, and N. Memon, "Pitch: Ai-assisted tagging of deepfake audio calls using challenge-response," arXiv preprint arXiv:2402.18085, 2024. [Online]. Available: https://arxiv.org/abs/2402.18085

[5] B. Zhang, H. Cui, V. Nguyen et al., "Audio deepfake detection: What has been achieved and what lies ahead," Sensors, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11991371/

[6] N. Chakravarty and M. Dua, "Improved feature extraction for hindi language audio impersonation attack detection," Multimedia Tools and Applications, 2024.

[7] R. Ranjan, L. Ayinala, M. Vatsa, and R. Singh, "Multimodal zero-shot framework for deepfake hate speech detection in low-resource languages," arXiv preprint arXiv:2506.08372, 2025. [Online]. Available: https://arxiv.org/abs/2506.08372

[8] A. Khan, K. M. Malik, J. Ryan et al., "Battling voice spoofing: A comparative analysis," Artificial Intelligence Review, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10462-023-10539-8

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

[10] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio deepfake approaches," IEEE Access, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10320354

[11] M. Li, Y. Ahmadiadli, and X. P. Zhang, "Audio anti-spoofing detection: A survey," arXiv preprint arXiv:2404.13914, 2024. [Online]. Available: https://arxiv.org/abs/2404.13914

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

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

[14] S. Saha, M. Sahidullah, and S. Das, "Exploring green ai for audio deepfake detection," arXiv preprint arXiv:2403.14290, 2024. [Online]. Available: https://arxiv.org/abs/2403.14290

[15] G. Tahaoglu, A. Gokceoglu, and A. Koivisto, "Deepfake audio detection with spectral features and resnext-based architecture," Expert Systems with Applications, vol. 235, p. 121293, 2025.

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

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

[18] R. Ranjan, M. Vatsa, and R. Singh, "Uncovering the deceptions: Analysis on audio spoofing detection," arXiv preprint arXiv:2307.06669, 2023. [Online]. Available: https://arxiv.org/abs/2307.06669

[19] C. O. Mawalim, Y. Wang, S. Okada, and M. Unoki, "Jmad: Multilingual audio deepfakes dataset for robust and generalizable detection," Preprint, 2025. [Online]. Available: https://candyolivia.github.io/assets/pdf/paper/JMADv1.pdf

[20] R. M. Purohit, A. J. Shah, D. H. Vaghera, and H. A. Patil, "Mladdc: Multi-lingual audio deepfake detection corpus," OpenReview, 2024. [Online]. Available: https://openreview.net/forum?id=ic3HvoOTeU

[21] A. Pianese, D. Cozzolino, G. Poggi et al., "Deepfake audio detection by speaker verification," in IEEE International Conference, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9975428

[22] S. Kaur, M. Buhari, N. Khandelwal, P. Tyagi, and K. Sharma, "Hindi audio-video-deepfake (hav-df): A hindi language-based audio-video deepfake dataset," 2024, school of Engineering & Technology, BML Munjal University, Gurugram, India. [Online]. Available: https://arxiv.org/abs/2411.15457

[23] Z. Wu, H. Delgado, M. Todisco, and N. Evans, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," in Proc. Interspeech, 2023. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2023/wu23_asvspoof.html

[24] A. R. Ambili and R. C. Roy, "Multi-tasking synthetic speech detection on indian languages," in IEEE International Conference on Signal Processing and Communications, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9744221/

[25] 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. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10650962/

[26] T. Tran and et al., "Parallelchain lab's anti-spoofing systems for asvspoof 5," in Proc. ASVspoof 2024, 2024. [Online]. Available: https://www.isca-archive.org/asvspoof_2024/tran24_asvspoof.pdf

[27] A. Javed, K. M. Malik, A. Irtaza et al., "Voice spoofing detector: A unified anti-spoofing framework," Expert Systems with Applications, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422002330

[28] M. Taeb, I. Kola-Adelakin, and H. Chi, "Forensic investigation of synthetic voice spoofing detection in social apps," in ACM Conference, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3696673.3723086

[29] D. Salvi, P. Bestagini, and S. Tubaro, "Synthetic speech detection through audio folding," in ACM Workshop, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3592572.3592844

[30] H. Tak, M. Todisco, X. Wang et al., "Automatic speaker verification spoofing using wav2vec 2.0," arXiv preprint arXiv:2202.12233, 2022. [Online]. Available: https://arxiv.org/abs/2202.12233

[31] V. Velumani, P. Sekar, M. Subramanian, and H. Mahaveer Chand, "Deepfake detection of images," ResearchGate Preprint, 2024. [Online]. Available: https://www.researchgate.net/publication/380854768_Deepfake_Detection_Of_Images

[32] N. M. Müller, P. Kawa, S. Hu et al., "A new approach to voice authenticity," arXiv preprint arXiv:2402.06304, 2024. [Online]. Available: https://arxiv.org/abs/2402.06304

[33] S. T. Yalla, M. G. P. Raju, and D. Nagaraju, "Decoding voice authenticity: Deep learning and audio features," in IEEE Conference, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10914906/

[34] O. C. Phukan, G. S. Kashyap, and A. B. Buduru, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," arXiv preprint arXiv:2404.00809, 2024. [Online]. Available: https://arxiv.org/abs/2404.00809

[35] N. Chakravarty and M. Dua, "Spectrogram-resnet41 for audio spoof attack detection with indian languages," Journal of System Assurance Engineering and Management, 2024.

[36] Y. Hou, H. Fu, C. Chen, Z. Li, H. Zhang, and J. Zhao, "Polyglotfake: A novel multilingual and multimodal deepfake dataset," in Proc. International Conference on Artificial Intelligence. Springer, 2024.

[37] A. Cohen, D. Shyrman, and A. Solonskyi, "Robust prosody modeling for synthetic speech detection," SSRN, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4892094

[38] R. M. Purohit, A. J. Shah, and 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. [Online]. Available: http://www.apsipa2024.org/files/papers/327.pdf

[39] R. Ranjan, B. Dutta, and M. Vatsa, "Faking fluent: Unveiling the achilles' heel of multilingual deepfake detection," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10744454/

[40] J. Fernandez, C. Lopez, and P. Garcia, "Benchmarking multilingual deepfake speech detectors on real-world scenarios," arXiv preprint arXiv:2409.08123, 2024. [Online]. Available: https://arxiv.org/abs/2409.08123

[41] R. Ranjan, K. Pipariya, M. Vatsa, and R. Singh, "Synhate: Detecting hate speech in synthetic deepfake audio in indic languages," arXiv preprint arXiv:2506.06772, 2025. [Online]. Available: https://arxiv.org/abs/2506.06772

[42] S. Sarala, M. Suresh Reddy, N. Sai Kiran Reddy, and V. Sai Sharan, "Deepfake detection on social media," International Journal for Research Trends and Innovation (IJRTI), vol. 9, no. 4, pp. 288–291, 2024. [Online]. Available: http://www.ijrti.org/papers/IJRTI2404040.pdf

[43] J. Guan, J. Li, and X. Chen, "A survey on speech deepfake detection: Taxonomy, challenges, and future directions," IEEE Access, 2023.

[44] H. Li, J. Wang, and T. Zhao, "Voiceguard: Robust detection of voice deepfakes with contrastive learning," in Proc. IEEE ICASSP, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10096543

[45] Y. Wu, X. Yang, and Z. Li, "Robust audio deepfake detection via multi-level spectrogram features," arXiv preprint arXiv:2401.08912, 2024. [Online]. Available: https://arxiv.org/abs/2401.08912

[46] A. Singh, P. Sharma, and R. Kumar, “Synthetic voice spoofing detection using hybrid cnn-lstm models,” Neural Computing and Applications, 2024.

[47] Y. Zhang, K. Qian, and H. Li, “Contrastive representation learning for generalizable audio deepfake detection,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10744488/

[48] A. Gupta, A. Kumar, and R. Singh, “Deepfake audio detection: A comprehensive review of challenges and countermeasures,” Multimedia Systems, 2024.

[49] L. Sun, J. He, and C. Wang, “Voicetrust: Reliable detection of audio deepfakes using phoneme-aware features,” in Proc. ACM Multimedia, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3581783.3612321

[50] M. Alshamrani, F. Khan, and H. Patel, “Deepfake speech detection: A systematic literature review,” arXiv preprint arXiv:2502.06712, 2025. [Online]. Available: https://arxiv.org/abs/2502.06712

[51] Y. Yang, X. Wang, and J. Liu, “Self-supervised learning for audio deepfake detection,” Pattern Recognition Letters, 2024.

[52] S. Cheng, Z. Yu, and Q. Li, “Specdefend: Detecting deepfake audio via spectral defenses,” in Proc. IEEE ICIP, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10239485

[53] J. Park, H. Lee, and S. Cho, “Few-shot audio deepfake detection with meta-learning,” arXiv preprint arXiv:2405.02345, 2024. [Online]. Available: https://arxiv.org/abs/2405.02345

[54] R. Mohanty, A. Patel, and N. Kumar, “Cross-lingual deepfake speech detection with transfer learning,” Speech Communication, 2024.

[55] V. Rao, A. Bhatia, and P. Singh, “Hybrid cnn-transformer networks for robust deepfake audio detection,” in Proc. IEEE International Joint Conference on Biometrics, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10899876

SANDEEP VERMA is currently working as an Assistant Professor-Research in the Department of Computer Science and Engineering at Thapar Institute of Engineering & Technology, Patiala, India. Dr. Verma received his Ph.D. from Dr BR Ambedkar National Institute of Technology Jalandhar, Punjab India in December 2020 in the ECE Department. His research interests include IoT, WSN, Intelligent Transportation System, Smart City, etc. Dr Verma has also a post doctoral experience from IIT Kharagpur, India wherein he worked for 5G standardization project. He has more than 30 publications in reputed journals and also serves as an Associate Editor of IET Wireless Sensor System Journal.

SEEMA BAWA Professor of Computer Science Engineering at Thapar Institute, holds an M.Tech. from IIT Kharagpur and a Ph.D. from TIET, Patiala, with over 26 years of academic and research experience. Her work spans Generative AI, Quantum Machine Learning, NLP, Blockchain, and Energy-Aware Computing, supported by a strong foundation in statistics from IIT Kanpur. She has supervised 28 Ph.D. scholars, led multiple national RD projects, and published over 200 papers. A proponent of holistic well-being, she actively promotes Sahaj Marg Meditation and the Art of Living.

SACHIN KANSAL is currently working as an Assistant Professor, at Department Computer Science and Engineering Department, at Thapar University. He obtained his B.Tech (CSE) from the IIMT Engineering College, Meerut, India, and an M.Tech from the Indian Institute of Information Technology, Allahabad, India. He received his PhD degree from the Indian Institute of Technology Delhi, New Delhi, India, in 2020. He has over 39 research papers, out of which 16 are in international journals and 23 are in International conferences.

SHIVANE KAPOOR is a Computer Engineering undergraduate at Thapar Institute of Engineering Technology with strong interests in Machine Learning, Computer Vision, and Artificial Intelligence. He has built real-world ML models for gesture-based control systems and customer churn prediction, gaining extensive hands-on experience in data preprocessing, model evaluation, feature engineering, and hyperparameter tuning. During his AI internship, he developed RNN and LSTM-based models for sequential data and designed autonomous agents using modern tools. He ranked in the top 5% in the Adobe GenSolve Hackathon 2024 and actively contributes to open-source projects while remaining involved in continuous learning and technical exploration.

KAUSTUBH SINGH is a Computer Engineering undergraduate at Thapar Institute of Engineering and Technology with strong interests in Machine Learning, Web Development, and AI-driven applications. He led the development of Project Sambhav, a socially driven platform built for an NGO to support unemployed individuals through career guidance and skill development. He also interned at Algorithm Research, Dubai, where he contributed to economic risk forecasting using machine learning, and has developed practical projects like Quickblog, an AI-assisted content creation platform. Kaustubh was selected among the top 250 nationwide for the JPMC Code for Good 2025 and ranked in the top 5% in the Adobe GenSolve Hackathon 2024. He is passionate about applying technology to solve real-world problems.

JAPNEET SINGH is an undergraduate Computer student with a strong foundation in Full Stack Web Development, Machine Learning, and Real-Time Systems. He specializes in the MERN stack and has built innovative applications like a smart expense tracker with OTP authentication, budget analytics, and monthly email summaries. He is also skilled in system-level programming and has developed a 3D Maze Game using OpenGL, incorporating features such as dynamic difficulty adjustment, lighting effects, and a minimap. With a solid grasp of secure backend development, RESTful APIs, and scalable database design, Japneet continues to expand his portfolio while actively engaging in open-source projects and collaborative technical learning.

DIWAKAR NARAYAN SOOD is a final-year undergraduate student in Computer Engineering at Thapar Institute of Engineering and Technology. He has a strong foundation in Full Stack Development, specializing in the MERN stack and DevOps, with a solid grasp of secure backend development, RESTful APIs, and scalable database design. He has built innovative applications such as a smart expense tracker featuring OTP authentication, budget analytics, and monthly email summaries. His expertise also extends to Machine Learning, Artificial Intelligence, and Real-Time Systems, reflecting his commitment to advancing intelligent systems and deep learning applications. Alongside his technical pursuits, he brings strong leadership, marketing, and organizational skills through active participation in university clubs and societies.

I s an undergraduate Computer Engineering student at Thapar Institute of Engineering and Technology specializing in AI, full-stack development, and system automation. Proficient in Python, Java, R, and C++, with expertise in Spring Boot, Angular, Flask, and Streamlit, he builds scalable, secure, and data-driven applications. His projects include a LangChain-powered YouTube Summarizer and a JWT-secured Cafe Management System. At Airtel Payments Bank and DisruptiveNext, he optimized fintech platforms—cutting API response times by 35%, scaling to 1M+ users, and accelerating AI deployment by 50%. A Best Data Science Expert Awardee and Google HACK4CHANGE Pre-finalist, he also serves as Director General Head of Hospitality for Thapar MUN Society, blending engineering rigor with leadership to solve real-world problems.