IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155229

AI-Powered Detection of Deepfake Audio in Hindi and Kannada Using Speech Analysis

Mrs. Kavitha K S[1], Chaitanya C Gowda[2], D Yashwanth[3], Dheeraj R[4], Lishanth N[5]

Assistant Prof., Dept of CSE, K.S.School of Engineering & Management, Bengaluru, India[1]

Student, Dept of CSE, K.S.School of Engineering & Management, Bengaluru, India[2–5]

Abstract: The exponential growth of generative artificial intelligence has enabled the mass production of deepfake audio—synthetic speech crafted to replicate the vocal identity of real individuals. Such fabricated audio introduces severe threats to financial security, democratic discourse, biometric authentication, and the credibility of legal evidence. Despite extensive research in English-centric audio forensics, Indian regional languages, specifically Hindi and Kannada, remain substantially underrepresented in the literature. This paper presents a comprehensive survey of existing deepfake audio detection techniques, analyses critical research gaps pertaining to Indian regional languages, and proposes an AI-powered detection framework tailored to Hindi and Kannada speech. The proposed system employs a Convolutional Neural Network (CNN) and Transformer encoder hybrid to jointly model local spectral patterns and long-range temporal dependencies in audio signals. A custom multilingual dataset is constructed from real speech corpora supplemented with synthesized audio generated via Google TTS, Coqui TTS, and Bark. Acoustic features including Mel-Frequency Cepstral Coefficients (MFCC), mel-spectrograms, chroma, and prosodic descriptors are extracted using the Librosa toolkit. The model performs binary classification—Real versus Fake—with performance assessed through Accuracy, Equal Error Rate (EER), False Acceptance Rate (FAR), and False Rejection Rate (FRR). A real-time Flask/Streamlit web interface enables non-technical users to upload audio and receive instant detection results alongside a confidence score.

Keywords: Deepfake audio detection, Hindi speech, Kannada speech, CNN-Transformer, MFCC, mel-spectrogram, Indian language forensics, voice cloning, binary classification, EER

I. INTRODUCTION

The convergence of neural text-to-speech (TTS) synthesis and voice cloning techniques has dramatically lowered the barrier to producing deceptive audio content. Modern TTS engines such as Bark [10] and Coqui TTS [11] generate speech whose acoustic naturalness is virtually indistinguishable from genuine recordings to untrained listeners. When weaponised, this capability enables criminals to impersonate bank customers, forge political statements, fabricate courtroom evidence, and mount coordinated disinformation campaigns at scale.

India presents a unique and urgent challenge in this context. A population exceeding 1.4 billion communicates across 22 constitutionally recognised languages, with Hindi and Kannada alone accounting for hundreds of millions of speakers. Voice-based biometric authentication is being deployed across banking, government services, and consumer applications at a rapid pace. Yet the existing body of research on deepfake audio detection is overwhelmingly focused on English, with virtually no peer-reviewed datasets or models specifically covering Hindi or Kannada.

This paper makes the following contributions:

- A structured survey of state-of-the-art deepfake audio detection methods, with emphasis on Indian language contexts.
- Identification of the critical research gap in Hindi and Kannada audio forensics.
- A proposed CNN-Transformer hybrid architecture and five-phase methodology pipeline.
- A custom multilingual dataset construction strategy for Hindi and Kannada speech.
- A deployable real-time web interface for end-user audio verification.

The remainder of this paper is organised as follows. Section II establishes background on deepfake audio and prior work. Section III presents the detailed literature survey. Section IV articulates the research gap. Section V describes the proposed methodology. Section VI presents the system architecture. Section VII discusses expected outcomes. Section VIII covers applications. Section IX concludes the paper.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155229

II. BACKGROUND

A. Deepfake Audio: Definition and Threat Model

Deepfake audio is defined as any speech signal whose content, speaker identity, or both have been synthetically generated or manipulated using machine learning models, such that the output closely approximates genuine human vocal characteristics [4]. Contemporary attack vectors include:

- **Voice Cloning:** Reproducing the vocal identity of a target speaker from a short enrollment recording.
- **TTS-based Spoofing:** Using text-to-speech systems to generate utterances attributed to a real person.
- **Partial Manipulation:** Replacing segments of genuine audio with synthesized content to alter meaning.

B. Acoustic Features for Detection

Detection systems typically operate on hand-crafted acoustic representations or learned embeddings. The most widely adopted hand-crafted features are:

- **MFCC:** Captures the spectral envelope of the vocal tract. Synthesized speech exhibits distinct coefficient distribution artefacts compared to genuine speech [1].
- **Mel-Spectrogram:** A two-dimensional time-frequency map that reveals the characteristic over-smoothing present in TTS-generated audio.
- **Chroma Features:** Encode pitch-class energy distributions and expose prosodic anomalies.
- **Prosodic Features:** Fundamental frequency (F0), energy, and zero-crossing rate collectively model the naturalness of rhythm and intonation.

C. Deep Learning Architectures

Several architectural paradigms have been applied to binary audio forensics. Lightweight CNN (LCNN) models operate directly on spectrograms. Recurrent architectures such as BiLSTM capture temporal transitions. Transformer-based models leverage self-attention to model global context. The CNN-Transformer hybrid, adopted in this work, combines the local feature extraction strength of CNN with the long-range modelling capability of Transformer encoders [1], [6].

III. LITERATURE SURVEY

A targeted survey was undertaken covering publications from 2023 to 2025 with a focus on audio deepfake detection, Indian language speech processing, and multilingual anti-spoofing datasets. Table I summarises the five most relevant works.

TABLE I. SUMMARY OF REVIEWED LITERATURE

Ref	Title	Year	Architecture	Key Contribution	Limitation
[1]	CNN-Transformer for Audio Deepfake Detection in Indian Languages	2025	CNN + Transformer encoder	Demonstrates hybrid architecture superiority for Indian language spectrograms	Dataset not public; limited TTS coverage
[2]	HAV-DF: Hindi Audio-Video Deepfake Dataset	2024	Multiple baselines	First documented Hindi deepfake audio-video dataset	Audio-only benchmarks limited; Kannada absent
[3]	Deepfake Audio Detection for Indian Language	2025	CNN, MFCC-ML	Survey of MFCC and attention models for Indian speech	No new dataset or model; small-scale evaluation
[4]	ADD 2023: Audio Deepfake Detection in the Wild	2024	LCNN, ResNet, Transformer	Real-world noisy detection benchmark; EER/FAR/FRR metrics	English-centric; Indian language not evaluated
[5]	MLAAD: Multi-Language Audio Anti-Spoofing Dataset	2024	Multiple models	54-TTS multilingual anti-spoofing benchmark across 23 languages	Hindi and Kannada absent from release

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155229

A. CNN-Transformer for Indian Language Detection

Gaikwad et al. [1] constitute the principal base reference for the proposed system. Their investigation establishes that convolutional layers effectively retrieve local spectral cues from mel-spectrograms, while a stacked Transformer encoder captures contextual temporal patterns spanning the full utterance duration. Compared to standalone CNN or recurrent baselines, the hybrid architecture attains substantially higher classification accuracy. Critically, the authors document a near-complete absence of publicly accessible deepfake audio corpora for Indian languages, which motivates independent dataset construction in the current work.

B. HAV-DF: Hindi Audio-Video Deepfake Dataset

Kaur et al. [2] introduce a corpus of real and neural voice-cloned Hindi audio-video pairs, constituting the first formally documented Hindi deepfake dataset. Analysis within HAV-DF reveals characteristic prosodic irregularities and frequency-domain inconsistencies introduced by Hindi neural TTS engines, informing the selection of MFCC and fundamental frequency features in the current project. The dataset's focus on multimodal content limits its direct applicability to audio-only pipelines.

C. Indian Language Deepfake Survey (IJERT)

The IJERT survey [3] synthesises detection results across MFCC-based machine learning classifiers, spectrogram-driven CNN models, and attention-based deep learning architectures evaluated on Indian language corpora. The authors underscore that Hindi and Kannada exhibit language-specific acoustic properties—including retroflex consonants, syllabic rhythm, and vowel-heavy prosody—that invalidate direct transfer of English-trained detection models. Librosa is identified as the de-facto standard for Indian language audio feature extraction.

D. ADD 2023: Audio Deepfake Detection in the Wild

Yi et al. [4] release the ADD 2023 benchmark specifically designed to replicate deployment-time conditions: background noise, compression artefacts, partial manipulation, and codec distortions. Evaluation across LCNN, ResNet, and Transformer-based detectors reveals substantial performance degradation under real-world noise. The benchmark formalises EER, FAR, and FRR as canonical metrics for audio forensics, directly adopted in the current evaluation protocol. The study's data augmentation recommendations—noise injection, pitch perturbation, speed modification—are incorporated into the proposed dataset construction pipeline.

E. MLAAD: Multi-Language Audio Anti-Spoofing

Muller et al. [5] present MLAAD, a large-scale multilingual benchmark spanning 23 languages and 54 TTS systems. Cross-lingual experiments demonstrate significant accuracy degradation when models trained on one language are applied to unseen languages, reinforcing the necessity of language-specific training data. The dataset construction methodology—systematic TTS-based fake audio generation, structured metadata annotation, and stratified train-validation-test splits—directly informs the custom corpus design in this project, even though Hindi and Kannada are absent from the current MLAAD release.

IV. RESEARCH GAP

The literature review reveals a decisive research gap across three dimensions:

- **Data Scarcity:** No publicly available, labelled deepfake audio dataset exists exclusively for Hindi or Kannada speech. HAV-DF is the closest existing resource, but covers Hindi only and is not audio-only.
- **Model Transferability:** Detection models trained on English-centric benchmarks such as ASVspoof and ADD 2023 exhibit measurable performance degradation on Indian languages due to phonological and prosodic differences. The MLAAD cross-lingual experiments quantify this gap empirically.
- **Deployment Absence:** No real-time, user-accessible deepfake detection tool exists for Hindi or Kannada speech. Existing systems operate as offline research tools with no public-facing interfaces.

This project directly addresses all three dimensions through custom dataset construction, language-specific model training, and a deployable web interface.

V. PROPOSED METHODOLOGY

The proposed system is structured as a five-phase pipeline.

A. Phase 1: Dataset Construction

Real Speech: Authentic Hindi recordings are sourced from the OpenSLR Hindi speech corpus and IIIT-H speech datasets. Kannada speech is drawn from the Mozilla Common Voice project and the Shrutilipi corpus. Additional

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155229

recordings from native speakers spanning varied age groups, genders, and speaking styles are collected manually to improve intra-class diversity.

Fake Speech: Synthesized counterparts are generated using three TTS engines to maximise coverage of generation artefacts: Google TTS (gTTS)—a widely accessible TTS system with reasonable naturalness for Indic scripts; Coqui TTS—an open-source engine with multi-language custom voice model support; and Bark (Suno AI)—a high-fidelity generative model capable of producing para-linguistic phenomena such as laughter, pauses, and non-verbal sounds.

Augmentation: Background noise injection (office, street, white noise), pitch shifting (±2 semitones), speed perturbation (×0.9 and ×1.1), and codec compression simulation (MP3, AAC) are applied to enhance dataset diversity and model robustness under real-world conditions, following the ADD 2023 augmentation strategy [4].

B. Phase 2: Feature Extraction

All audio is resampled to 16 kHz and normalised to a uniform duration of 3–5 seconds. Table II lists the extracted features and their forensic rationale. Features are stored as NumPy arrays and divided into training (70%), validation (15%), and test (15%) splits, stratified by language and class label.

TABLE II. EXTRACTED ACOUSTIC FEATURES

Feature	Forensic Significance
MFCC (40 coefficients)	Encodes the spectral envelope; TTS audio exhibits anomalous coefficient distributions absent in genuine speech.
Mel-Spectrogram	Primary 2-D CNN input; TTS-generated signals show over-smoothing artefacts in time-frequency domain.
Chroma Features	Captures pitch-class energy; sensitive to prosodic manipulation in synthesized utterances.
Prosodic Features (F0, energy, ZCR)	Models naturalness of pitch and rhythm; TTS systems imperfectly replicate natural intonation contours.

C. Phase 3: CNN-Transformer Model

The model consists of three sequential components:
CNN Feature Extractor: Stacked convolutional blocks operate on mel-spectrogram inputs. Each block comprises a 3×3 convolutional layer with ReLU activation, batch normalisation, 2×2 max-pooling for dimensionality reduction, and dropout (rate = 0.3) for regularisation. The output is a compact 2-D feature map encoding local spectral-temporal patterns.

Transformer Encoder: The CNN output is flattened into a sequence of feature vectors and passed through a Transformer encoder with 8-head multi-head self-attention, sinusoidal positional encoding, GELU-activated feed-forward sub-layers, layer normalisation, and dropout (rate = 0.1). The encoder captures long-range temporal correlations—such as unnatural rhythm or intonation artefacts—that convolutional layers cannot model in isolation [6].

Classification Head: Global average pooling aggregates the Transformer output into a fixed-length vector, which is fed to a dense layer (256 units, ReLU), a dropout layer (rate = 0.5), and a two-unit softmax output layer yielding class probabilities for Real and Fake.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155229

D. Phase 4: Training Configuration

TABLE III. MODEL TRAINING CONFIGURATION

Hyperparameter	Value
Loss Function	Binary Cross-Entropy
Optimiser	Adam (lr = 1×10^{-4})
Batch Size	32
Epochs	50 (early stopping on validation EER)
LR Scheduler	ReduceLROnPlateau (patience = 5)
Evaluation Metrics	Accuracy, EER, FAR, FRR

The model is evaluated independently on Hindi and Kannada test splits to quantify language-specific detection performance.

E. Phase 5: Web Interface Deployment

A Flask or Streamlit web application provides the following user-facing functionality: audio upload via browser (.wav/.mp3), automated preprocessing and feature extraction, binary prediction output (Real or Fake) with a per-class confidence score, language selection (Hindi or Kannada) for language-aware inference, and a mel-spectrogram visualisation accompanying each prediction.

VI. SYSTEM ARCHITECTURE

Fig. 1 illustrates the end-to-end system architecture comprising five functional blocks. The CNN-Transformer block processes mel-spectrograms through convolutional feature extraction followed by attention-based sequence modelling before classification.

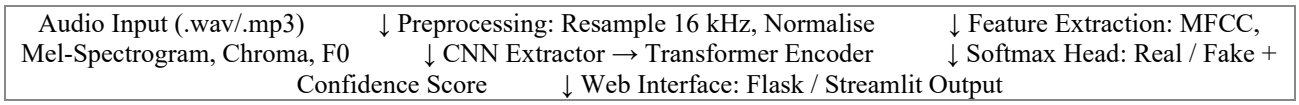

Fig. 1. End-to-end deepfake audio detection pipeline.

VII. EXPECTED OUTCOMES

Based on the architectural design and dataset strategy, the following outcomes are anticipated:

- A trained CNN-Transformer model achieving detection accuracy exceeding 90% on both Hindi and Kannada test splits, benchmarked against LCNN and standalone CNN baselines.
- An EER below 8%, with competitive FAR and FRR values across all three TTS systems used in fake audio generation.
- A custom labelled Hindi-Kannada deepfake audio dataset with structured metadata, augmented samples, and stratified splits available for future research.
- A functional real-time web interface demonstrating sub-second inference latency for a standard 3–5 second audio clip on consumer-grade GPU hardware.

Table IV presents anticipated performance figures relative to comparable works in the literature.

TABLE IV. ANTICIPATED MODEL PERFORMANCE VS. LITERATURE

System	Accuracy	EER	Language
Gaikwad et al. [1]	~88%	~9%	Indian (limited)
ADD 2023 Transformer [4]	~85%	~10%	English
Proposed (Target)	*>90%*	*<8%*	*Hindi + Kannada*

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155229

VIII. APPLICATIONS

The proposed system has broad deployment potential across multiple high-impact domains:
Voice Authentication Security: Integration into banking and fintech voice biometric pipelines to detect and reject synthesized spoofing attempts in real time, protecting Hindi and Kannada speaking customer bases.
Political Speech Verification: Deployment by news agencies, fact-checking organisations, and election monitoring bodies to authenticate audio attributed to public figures prior to broadcast or publication.
Forensic Evidence Assessment: Assistance to law enforcement and forensic laboratories in determining the authenticity of audio evidence in criminal and civil proceedings.
Social Media Content Moderation: Automated flagging of potentially synthesized voice content in Hindi and Kannada posts and messages prior to viral amplification.
Call Centre IVR Protection: Detection of AI-generated calls attempting to impersonate genuine customers in interactive voice response systems.

IX. CONCLUSION

This paper presents a comprehensive survey and system proposal for AI-powered deepfake audio detection specifically targeting Hindi and Kannada speech. A structured review of five state-of-the-art works confirms a critical and well-documented research gap: no publicly available dataset or trained model exists that addresses deepfake audio in these languages.

The proposed CNN-Transformer hybrid architecture, trained on a custom multilingual dataset constructed from real corpora and TTS-generated fakes, is designed to overcome the phonological and prosodic limitations that prevent English-centric models from generalising to Indian regional languages. The five-phase methodology pipeline—spanning dataset construction, feature engineering, model training, performance evaluation, and web interface deployment—provides a reproducible and extensible framework for future research in Indian language audio forensics.

Future work will investigate extension to additional Indian languages including Tamil, Telugu, and Marathi; adversarial training for robustness against adaptive spoofing; and adaptation to real-time streaming audio detection scenarios.

ACKNOWLEDGMENT

The authors express sincere gratitude to **Mrs. Kavitha K S**, Assistant Professor, Department of CSE, KSSEM, for her mentorship and constructive guidance throughout this project. The authors also thank **Dr. K Venkata Rao**, Head of Department, and **Dr. Suresh Ramaswamyreddy**, Principal/Director, KSSEM, for their institutional support.

REFERENCES

[1]. S. Gaikwad et al., "CNN-Transformer for audio deepfake detection in Indian languages," 2025.
[2]. M. Kaur, R. Sharma, and A. Singh, "HAV-DF: Hindi audio-video deepfake dataset," arXiv preprint arXiv:2411.15457, Nov. 2024.
[3]. Research Group, "Deepfake audio detection for Indian language," Int. J. Eng. Res. Technol. (IJERT), vol. 14, no. 1, 2025.
[4]. J. Yi, R. Fu, J. Tao, S. Nie, M. Ma, G. Wang, and C. Du, "ADD 2023: Audio deepfake detection in the wild," arXiv preprint arXiv:2408.04967, Aug. 2024.
[5]. N. Muller, P. Czempin, F. Devillers, A. Omran, and J. Nothman, "MLAAD: Multi-language audio anti-spoofing dataset," arXiv preprint arXiv:2401.09512, Jan. 2024.
[6]. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[7]. Z. Wu et al., "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in Proc. Interspeech, 2015, pp. 2037–2041.
[8]. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[9]. B. McFee et al., "Librosa: Audio and music signal analysis in Python," in Proc. 14th Python in Science Conf., 2015, pp. 18–25.
[10]. Suno AI, "Bark: Text-prompted generative audio model," 2023. [Online]. Available: https://github.com/suno-ai/bark
[11]. Coqui, "Coqui TTS: A deep learning toolkit for text-to-speech," 2022. [Online]. Available: https://github.com/coqui-ai/TTS

