Explainable Audio Deepfake Detection

Akanksha Singh Geetha Raju Gokul S Krishnan B Ravindran

Centre for Responsible AI, Wadhwani School of Data Science and AI, IIT Madras

Contact: akanksha@cerai.in

Introduction

Motivation: Discussions with subject experts revealed that fact-checkers in India face growing challenges from audio deepfakes rather than visual ones. Audio spoofs are subtler, often masked by background noise, music, and compression artifacts, making detection for both humans and AI more difficult. Additionally, Indian languages remain underrepresented compared to English and other global languages.
Goal: Develop an *explainable* audio deepfake detection framework that grounds its decisions in phonetic and morphological cues.
Contributions: **(1)** Benchmarking five detection models across nine diverse datasets (English, Hindi, and multilingual). **(2)** Systematic evaluation using eight metrics, including EER and AUPRC.

Data

Datasets: Models are trained on ASVspoof2019-LA, a benchmark dataset for audio deepfake detection. The training set contains 2,580 real and 22,800 fake samples, while the development set has 2,548 real and 22,296 fake samples.

Test Set	Real	Fake
ASVspoof2019-LA	7355	64578
ASVspoof2021-LA	18452	163114
ASVspoof2021-DF	22617	589212
Fake-or-Real	2264	2370
In The Wild	19963	11816
PartialSpoof	7355	63882
IndieFake	8172	11342
MLAAD	-	201000
HAV-DF	196	322

Table 1. Distribution of real and fake samples across datasets.

Characteristics: These datasets range from state-of-the-art benchmarks (ASVspoof) to real-world noisy data (In The Wild), partially spoofed subsets (PartialSpoof), and multilingual corpora (MLAAD), including Indic languages such as Hindi, Marathi, Tamil, and Bengali. Spoof types encompass GAN-based, diffusion-based, and internet-sourced synthetic speech.

Methodology

Models:

- **RawNet2** Tak, Patino, et al. 2021: End-to-end raw waveform model capturing fine-grained artifacts often missed by handcrafted or spectrogram features.
- **RawGAT-ST** Tak, Jung, et al. 2021: Spectro-temporal graph attention network that jointly models spectral and temporal dependencies to capture cross-domain artifacts.
- **AASIST** Jung et al. 2021: Integrates heterogeneous spectro-temporal attention for unified artifact representation.
- **SSL W2V2** Tak, Todisco, et al. 2022: Self-supervised framework learning robust, transferable speech representations.
- **Conformer** Rosello et al. 2023: Combines convolutional and transformer blocks for local and global context modelling in speech.

Metrics:

- **Accuracy:** Reports overall classification performance and per-class accuracies for real and fake samples.
- **AUROC and EER:** The *Area Under the ROC Curve* and *Equal Error Rate* measure how well a model separates real and fake samples across thresholds. AUROC reflects threshold-independent discrimination, while EER indicates the balance point between false positives and false negatives—lower EER denotes better calibration.
- **AUPRC and TPR@90%FPR:** The *Area Under the Precision–Recall Curve* is more robust under class imbalance, focusing on the correct detection of the minority (fake) class. *True Positive Rate at 90% False Positive Rate* evaluates sensitivity under strict false alarm constraints.

Results

In-Domain Performance: Models achieve high accuracy on benchmark datasets (ASVspoof2019/2021), with SSL-W2V2 and Conformer showing consistently strong performance in both binary and class-wise evaluations.
Out-of-Domain Generalization: Performance drops on unseen datasets such as Fake-or-Real and In-The-Wild, more so in the Real class than the Fake class. The PartialSpoof Fake samples are also harder to detect, whereas the Real samples remain largely stable.

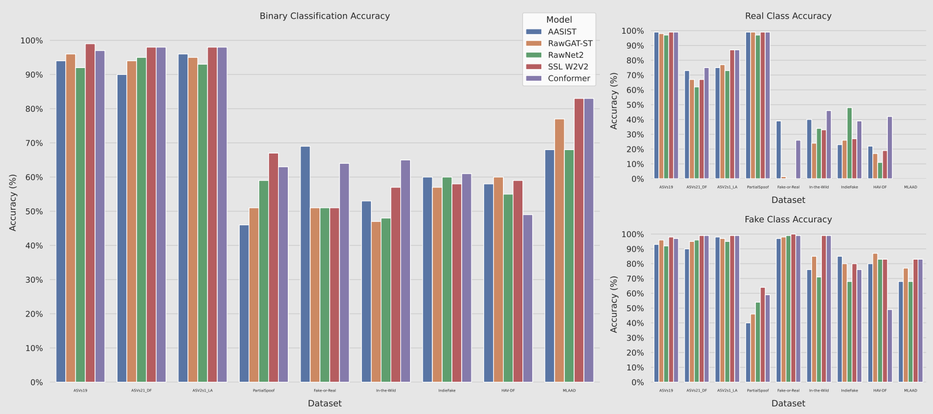

Figure 1. Model performance across datasets.

Multilingual Benchmarking: THE SSL W2V2 model consistently outperforms other models across Indian languages. Accuracy estimates for languages with fewer samples may be less reliable, but overall trends indicate robust multilingual performance.

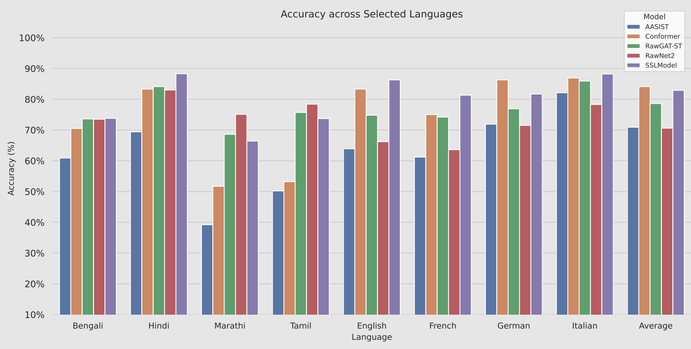

Figure 2. Multilingual evaluation on the MLAAD dataset.

Evaluation

Key Insights:

- Model performance varies consistently across datasets, highlighting domain sensitivity.
- Partial spoofs remain challenging to detect compared to fully synthetic audio.
- Models capture fake characteristics better than real speech features.
- Current models are largely language-agnostic but still struggle with low-resource languages.
- Metrics such as AUROC and EER are useful in academic settings, while AUPRC and TPR@90%FPR are more interpretable for real-world deployment.

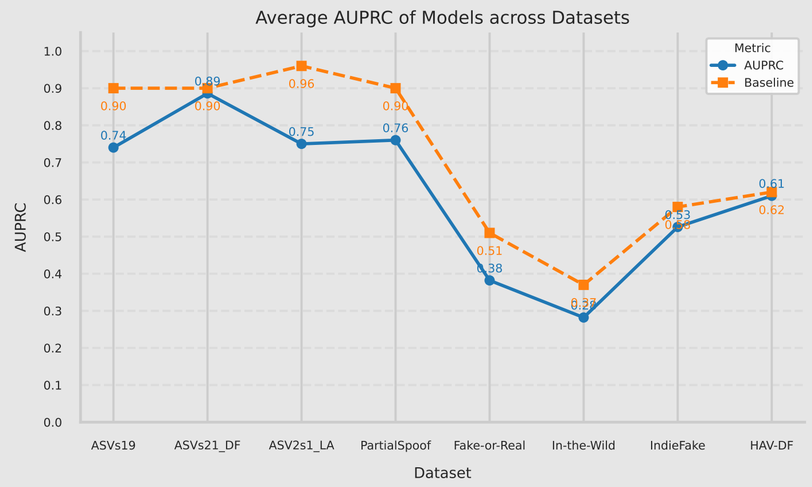

Figure 3. AUPRC comparison across datasets.

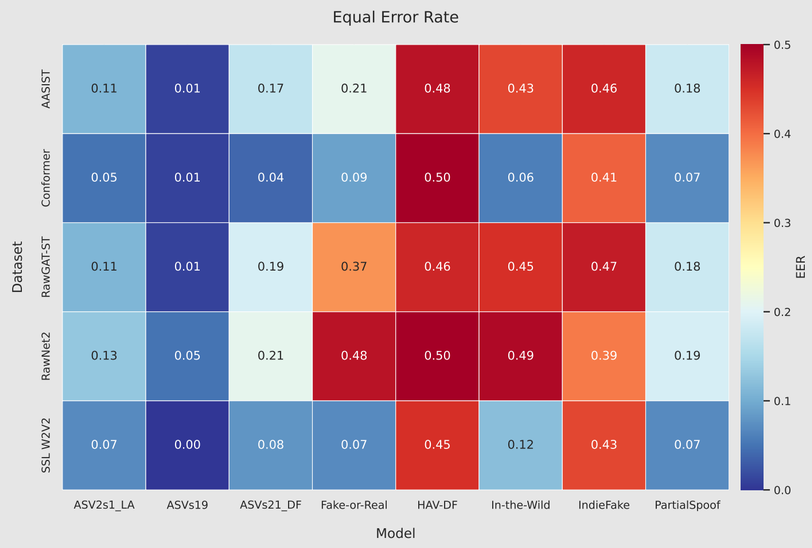

Figure 4. EER comparison across models.

Limitations: Despite strong in-domain accuracy, models show limited understanding of real vs. fake speech, poor cross-dataset generalisation, and reduced sensitivity to language-specific cues. Explainable models by design are needed to bridge these gaps.

Conclusion & Future Work

Conclusion. This study benchmarks state-of-the-art models for audio deepfake detection under a linguistically grounded, explainable paradigm. Transformer-based models (Conformer, SSL W2V2) excel in metrics but still lack explainability and cross-lingual generalisation.
Future Directions.

- Develop a multilingual corpus with naturalistic speech and varied synthesis pipelines.
- Integrate phoneme-aware embeddings and interpretable latent features.
- Build an open-source fact-checking plugin offering real-time, explainable predictions.

References

Rosello, E., A. Gomez-Alanis, A. M. Gomez, and A. Peinado (2023). "A conformer-based classifier for variable-length utterance processing in anti-spoofing". In: *Interspeech 2023*, pp. 5281–5285. DOI: 10.21437/Interspeech.2023-1820.

Tak, H., M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans (2022). *Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation*. arXiv: 2202.12233 [eess.AS]. URL: https://arxiv.org/abs/2202.12233.

Jung, J.-w., H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans (2021). *AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks*. arXiv: 2110.01200 [eess.AS]. URL: https://arxiv.org/abs/2110.01200.

Tak, H., J.-w. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans (2021). *End-to-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection*. arXiv: 2107.12710 [eess.AS]. URL: https://arxiv.org/abs/2107.12710.

Tak, H., J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher (2021). *End-to-end anti-spoofing with RawNet2*. arXiv: 2011.01108 [eess.AS]. URL: https://arxiv.org/abs/2011.01108.