

# Wav2Vec Meets Conformer: A Novel Hybrid Approach for Multilingual Deepfake Audio Detection

USHA JANAKIRAMAN<sup>1</sup>, PRIYADHARSHINI AMBALAVANAN<sup>2</sup>, PADMAPRIYA S<sup>3</sup>

<sup>1,2</sup>Student, Artificial Intelligence and Data Science, Panimalar Engineering College - Chennai

<sup>3</sup>Assistant Professor, Artificial Intelligence and Data Science, Panimalar Engineering College - Chennai

**Abstract:** Deepfake audio refers to synthetic speech that closely mimics a person's voice, posing risks to security and privacy. This paper proposes a hybrid detection framework combining XLS-R, a multilingual speech representation model, with the Conformer architecture, which captures both local and global audio dependencies. XLS-R extracts rich multilingual embeddings, while the Conformer leverages temporal and contextual features to distinguish genuine from AI-generated speech. Evaluation on benchmark datasets demonstrates that the proposed system achieves improved accuracy and robustness across multiple languages and acoustic conditions.

**Keywords:** Conformer, Deepfake Audio, Multilingual Speech Representation, XLS-R

## I. INTRODUCTION

Deepfake audio refers to synthetic or manipulated speech that closely mimics a person's voice, generated using advanced AI techniques. With rapid advancements in speech synthesis, these audio forgeries pose significant risks to privacy, security, and trust in digital communication. Malicious actors can exploit deepfake audio for impersonation, spreading misinformation, fraud, or identity theft, making the development of reliable detection mechanisms increasingly important.

Existing deepfake audio detection methods, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or spectrogram-based analysis, have shown promising results in controlled environments. However, many of these methods struggle to generalize across multilingual datasets, different speakers, and diverse acoustic conditions. Additionally, most approaches rely on handcrafted features or are optimized for specific datasets, limiting their applicability in real-world scenarios. These challenges highlight the need for robust and scalable detection frameworks capable of cross-lingual and cross-environment performance.

To address these limitations, this work proposes a hybrid detection framework that integrates XLS-R, a self-supervised multilingual speech representation model, with the Conformer architecture. XLS-R extracts rich embeddings that capture subtle variations in speech across multiple languages, while the Conformer effectively models temporal and contextual dependencies within audio sequences. This combination allows the system to leverage both global and local audio patterns, improving its ability to distinguish genuine speech from AI-generated content.

The hybrid approach merges multilingual feature extraction with context-aware sequential modeling, providing a unified detection pipeline. By combining these complementary strengths, the framework enhances accuracy, robustness, and generalization, addressing the shortcomings of prior methods. This design makes it suitable for practical applications, including forensic investigations, social media monitoring, and verification of user-generated audio content.

The proposed system is evaluated on the MLAAD benchmark dataset, which contains genuine and synthetic audio samples across 23 languages. Experimental results demonstrate that the framework outperforms existing baselines in accuracy, robustness, and cross-lingual generalization, providing a reliable and effective solution for deepfake audio detection across diverse languages and acoustic conditions.

## II. LITERATURE SURVEY

Deepfake audio detection has become a significant and urgent challenge in digital security due to the increasing sophistication of AI-generated speech, which can closely mimic real human voices [1]. The rapid advancement of deep learning and speech synthesis technologies has made it possible to

produce highly realistic audio, raising concerns about fraud, misinformation, and identity theft. Early detection methods primarily relied on convolutional neural networks (CNNs) applied to audio spectrograms. These methods effectively captured subtle spectral and temporal patterns that differ between genuine and synthetic speech. Although promising results were achieved in terms of accuracy and balanced precision-recall scores, some misclassifications persisted, motivating further exploration of preprocessing techniques, data augmentation, and more advanced network architectures [1].

Single-modality approaches, which rely solely on either raw waveform signals or spectrogram representations, often face challenges under real-world conditions. Factors such as background noise, variable acoustic environments, and unseen forgery techniques reduce their robustness and generalization ability [2]. To overcome these limitations, cross-modal and multi-scale approaches have been developed. By integrating complementary information from both waveform and spectrogram representations, these methods capture discriminative artifacts more effectively. For example, the WaveSpec framework leverages Wav2Vec2 for encoding raw waveforms and a UNet-based architecture for spectrogram analysis. A Cross-Modal Feature Fusion module then combines these representations, improving the system's robustness against previously unseen deepfakes. Additional constraints on embeddings of genuine audio further enhance detection performance for out-of-domain samples [2].

Another significant approach involves the use of Teager Energy Cepstral Coefficients (TECC), which capture nonlinear energy fluctuations in speech. When combined with a ResNet-50 classifier, TECC features outperform traditional MFCC and LFCC representations, demonstrating superior resistance to noise and better discrimination between real and synthetic audio signals [3]. Alongside these feature-based methods, real-time detection techniques combine deep learning with signal processing to analyze voice patterns and identify anomalies rapidly. These approaches are particularly useful in practical applications, such as forensic audio analysis, where low-latency, high-accuracy detection is critical [4].

Comprehensive reviews of deepfake audio detection methods reveal a wide spectrum of strategies. Traditional machine learning models—including Support Vector Machines (SVMs) and Decision Trees (DTs)—as well as deep learning architectures such as CNNs, RNNs, and hybrid models, have been extensively applied. SVM-based approaches have reported accuracies as high as 99%, whereas decision trees achieved around 73% accuracy. Advanced evaluation metrics, including Equal Error Rate (EER) and tandem Detection Cost Function (t-DCF), are often used to better quantify model performance. Notably, Siamese CNN models have shown substantial improvements over baseline methods by learning more robust representations for detecting subtle differences between real and synthetic audio [5].

Hybrid models further enhance detection performance. The MFCC-GNB XtractNet framework, for example, combines MFCC feature extraction with Gaussian Naive Bayes (GNB) and Non-Negative Matrix Factorization. This approach achieves near-perfect detection accuracy (~99.93%) and demonstrates the effectiveness of probabilistic feature transformation combined with machine learning integration for robust audio authentication [6]. Similarly, CNN-Transformer hybrid architectures have been proposed for multilingual deepfake detection in Indian languages such as Hindi, Marathi, Tamil, Telugu, Malayalam, and Kannada. These models leverage CNNs for spatial feature extraction and Transformer layers to capture temporal dependencies, integrating multiple audio representations (Mel spectrograms, LFCC, and Wav2Vec embeddings) through spectral-temporal gating mechanisms. Optimizations such as pretraining, knowledge distillation, quantization, and pruning enable high generalization and real-time deployment on edge devices [7].

Biologically inspired approaches also show promising results. The BTS-E (Breathing-Talking-Silence Encoder) framework utilizes human breathing patterns, which are difficult for text-to-speech systems to replicate. By extracting frame-level LFCC features and classifying segments into breathing, talking, or silence using Gaussian Mixture Models, a Segmentation Encoding Vector is formed. This vector is then used in a positional correlation encoding network to enhance a baseline countermeasure system (e.g., RawNet2). Evaluations

on ASVspoof 2019 and 2021 datasets show that incorporating breathing cues can improve detection accuracy by up to 46%, demonstrating the potential of leveraging intrinsic human-generated sounds for deepfake detection [8].

Multi-stage frameworks enhance detection further by using pre-trained models to extract phonetic, speaker identity, and prosodic features, combined with GAN-based synthetic data augmentation. Contrastive learning is incorporated to improve feature discrimination, strengthening the model's ability to distinguish real and synthetic audio. These strategies provide enhanced robustness and generalization compared to conventional single-stage models [9].

Audio dub detection has emerged as another effective approach, using multi-modal AI to detect AI-generated audio clips in both real-time and offline settings. Hybrid networks combining CNNs for spatial features and BiLSTM layers for temporal dependencies, alongside incremental learning frameworks like Adaptive Random Forest, allow efficient streaming and reduce computational costs. Spectral and temporal feature integration further improves generalization and robustness, making these systems suitable for deployment in live audio monitoring and forensic investigations [10].

Fine-grained detection of partially spoofed audio segments has also been explored. The Temporal Deepfake Location (TDL) method separates real and fake audio frames using an embedding similarity module and applies temporal convolution to capture positional information. Evaluations on the ASVspoof 2019 Partial Spoof dataset show that TDL accurately localizes partially spoofed audio while maintaining cross-dataset generalization, highlighting its utility for detecting subtle manipulations [11].

Unsupervised anomaly detection methods offer advantages in identifying novel deepfake audio. By learning the distribution of real human speech using mel-spectrogram and MFCC features, GAN-based models such as GANomaly and f-AnoGAN compute anomaly scores for detection without requiring labeled data. This enables real-time identification of previously unseen deepfakes. Optimal configurations, such as using a mel-spectrogram with 40 mel filters in GANomaly, achieve an F1-score of 0.93, demonstrating efficacy for applications like

voice phishing prevention, identity verification, and misinformation control [12].

Finally, audio splicing detection addresses cases where multiple speech segments from different recordings are concatenated to create fake audio. CNN-based methods extract device-specific embeddings, which are clustered using K-Means to detect splicing. Distance-based techniques and iterative multi-shift operations allow precise localization of splicing points. Evaluations on the MOBIPHONE dataset demonstrate 96% detection accuracy and a maximum localization error of only 0.012 seconds. This approach highlights the value of leveraging acquisition device traces in forensic audio analysis and lays the groundwork for future work, including detection of splicing in synthetic or transmitted audio [13].

Integrating multi-modal, multi-scale, and feature-transformed approaches significantly enhances the robustness and generalization of deepfake audio detection systems. Methods such as WaveSpec, TECC, real-time anomaly detection, hybrid MFCC-GNB frameworks, BTS-E, and TDL collectively provide resilient, high-accuracy systems capable of handling diverse languages, acoustic environments, and advanced manipulation techniques.

In comparison to the reviewed methods, the proposed hybrid approach demonstrates superior performance. By integrating XLS-R for rich multilingual embeddings with the Conformer architecture to capture both local and global audio dependencies, our model achieves robust, accurate, and real-time deepfake audio detection across multiple languages and acoustic conditions. This unified, multi-modal framework not only detects unseen and emerging deepfakes but also outperforms prior approaches in generalization, precision, and practical applicability.

### III. PROPOSED METHODOLOGY

The proposed work introduces a hybrid XLS-R + Conformer framework for multilingual audio deepfake detection. The design integrates cross-lingual self-supervised embeddings with temporal-context modeling to address challenges in detecting synthesized or manipulated speech under realistic conditions. This section details the data collection, preprocessing pipeline, model design, implementation details, and evaluation metrics.

### *A. Data Collection*

The system utilizes the Multilingual Audio Anti-Deepfake (MLAAD) dataset for both training and evaluation, as it is a large-scale benchmark specifically created for detecting audio forgeries. This dataset covers 23 languages, allowing the model to achieve strong cross-lingual generalization instead of being limited to a single language. It also includes a wide range of manipulation techniques such as Text-to-Speech (TTS) for AI-generated synthetic voices, Voice Conversion (VC) for style transfer while retaining the original linguistic content, and splicing or editing operations commonly used to modify speech continuity. In addition, the recordings span clean, noisy, and compressed environments, simulating real-world playback conditions. Each audio sample is thoroughly annotated with a class label indicating whether it is genuine or fake/cloned, along with detailed metadata describing the manipulation type, language, recording setup, and noise characteristics.

To address class imbalance often found in anti-spoofing datasets, augmentation techniques were applied to create a more balanced and diverse training set. Pitch shifting by  $\pm 2$  semitones was used to simulate variations in speakers, while time stretching within the range of  $0.9\text{--}1.1\times$  introduced natural differences in speech rhythm. Furthermore, noise injection—using both Gaussian noise and real-world noise samples drawn from the MUSAN dataset—helped replicate offline distortions and unpredictable background conditions. These augmentation strategies significantly improved the model's robustness, enabling better generalization across different languages, speaker variations, and environmental disturbances.

### *B. Pre-processing*

Preprocessing plays a crucial role in standardizing raw offline audio recordings before embedding extraction. The pipeline begins with noise reduction, where spectral subtraction and Wiener filtering are applied to suppress ambient disturbances while preserving subtle speech artifacts that deepfake systems typically fail to mimic. All audio is then resampled to 16 kHz mono WAV format, ensuring consistency and compatibility with the XLS-R pretrained feature extractor. Silence trimming is performed using an energy-based Voice Activity

Detector (VAD) to remove long non-speech segments, which helps avoid unnecessary computation and prevents the model from overfitting to silence. Additionally, amplitude normalization is applied to maintain loudness within a fixed range, ensuring that the model attends to spectral and temporal patterns rather than loudness fluctuations. To further enhance robustness, on-the-fly augmentation techniques such as time stretching and pitch shifting are integrated during training epochs, improving generalization and reducing overfitting. Overall, unlike traditional handcrafted features such as MFCC or LPC, this preprocessing pipeline effectively cleans and standardizes the input audio while retaining essential synthetic artifacts necessary for accurate deepfake detection.

### *C. Proposed Model*

The proposed XLS-R + Conformer hybrid model integrates multilingual embeddings with temporal-convolutional attention mechanisms to achieve highly robust deepfake voice detection. The first stage involves feature extraction using XLS-R, a wav2vec 2.0-based multilingual self-supervised model trained on 436,000 hours of speech across 128 languages. XLS-R provides rich embeddings that capture phonetic cues related to linguistic correctness, prosodic patterns such as tone, pitch, and rhythm, and speaker-specific characteristics like timbre and voice quality. Its cross-lingual capability significantly reduces retraining costs when working with new or low-resource languages, making it well-suited for universal deepfake detection.

In the second stage, contextual temporal modeling is performed using the Conformer architecture, which blends convolutional layers with self-attention mechanisms. While the convolutional layers detect local spectral and temporal variations, the self-attention layers capture broader speech context over long durations. This dual mechanism enables the model to identify subtle inconsistencies introduced by synthesis algorithms, such as unnatural pauses, pitch instability, or phase distortions—artifacts that commonly appear in AI-generated or voice-converted speech.

The outputs from the XLS-R feature extractor and the Conformer network are then fused to form a multi-level feature representation. This fusion preserves both local micro-patterns, including glitches or

unnatural harmonics, and global fluency patterns related to rhythm and prosody. Finally, a fully connected dense layer with Softmax activation produces the classification output, labeling each audio sample as either genuine (Class 0) or fake/cloned (Class 1). By combining multilingual speech representations with advanced temporal modeling, this architecture outperforms CNN-only or RNN-only approaches and achieves strong generalization across unseen languages and diverse spoofing attacks.

#### *D. Implementation*

The model is implemented using PyTorch, with HuggingFace Transformers handling the XLS-R and Conformer components. Training is carried out using the Adam optimizer with weight decay, and a learning rate of  $1e-4$  combined with a warmup phase followed by a ReduceLROnPlateau scheduler to stabilize convergence. A batch size of 32 is used with balanced mini-batches to ensure equal representation of genuine and fake samples, and training is conducted for 20 epochs with early stopping based on validation loss. To handle class imbalance, a Weighted CrossEntropyLoss function is applied, ensuring the model does not become biased toward either class.

The hardware setup includes training on an NVIDIA RTX 3060 GPU with full CUDA acceleration, while evaluation is performed on a CPU environment to simulate real-world offline use cases such as deployment on laptops or embedded systems. The software stack consists of Python 3.10, PyTorch 2.1, and HuggingFace Transformers 4.x, with Librosa used for audio preprocessing tasks and torchaudio employed for feature transformations. This combination ensures an efficient, scalable, and deployment-friendly training and evaluation pipeline.

#### *E. Evaluation Metrics*

The performance of the proposed XLS-R + Conformer hybrid model was thoroughly evaluated using standard classification metrics, demonstrating strong robustness across multilingual and noisy audio conditions. The model achieved impressive accuracy, with 98.45% on the test set and 98.91% on the validation set, indicating highly reliable classification of both real and fake audio samples.

Precision scores were consistently high, with Class 0 (real audio) reaching 98% on the test set and 99% on the validation set, while Class 1 (fake audio) achieved 99% in testing and 98% in validation. These results highlight the model's exceptional ability to correctly identify genuine and spoofed samples without producing misleading positive predictions. Recall values were similarly strong, with both real and fake audio classes scoring between 98–99% across test and validation datasets, demonstrating that the model effectively minimizes false negatives. Correspondingly, F1-scores reflected this balance, with both classes achieving 98–99%, confirming the model's optimal alignment of precision and recall across multilingual contexts and varied noise conditions.

Macro-average and weighted-average metrics further reinforce the model's balanced performance. Macro-average precision, recall, and F1-scores ranged from 98–98.5% across both datasets, while weighted averages remained slightly higher at around 98.4–98.8%, accounting for class distribution. These results show that the model maintains consistent performance even under class imbalance scenarios. The evaluation was conducted on a subset of the MLAAD dataset consisting of 4,500 samples each for the test and validation sets, with an equal split of 2,250 real and 2,250 fake audio clips. This distribution, derived from the MLAAD dataset's official split, ensures that the evaluation aligns with best practices for testing deepfake detection systems. Overall, the consistently high metrics across accuracy, precision, recall, and F1-score affirm that the proposed hybrid model delivers reliable, language-agnostic, and distortion-resilient deepfake detection.

#### *F. Insights*

- **High Precision and Accuracy:** The XLS-R + Conformer model is highly reliable in identifying both real and fake audio, achieving near-perfect accuracy on both test (98.45%) and validation (98.91%) sets.
- **Robust Detection Across Classes:** With recalls of 98% for both real and fake audio on the test set, and 99% for real audio and 98% for fake audio on the validation set, the model minimizes false negatives effectively.
- **Balanced Performance:** The F1-scores confirm consistent and balanced performance across

both real and fake audio classes, highlighting minimal trade-offs.

- **Cross-Lingual Robustness:** The integration of XLS-R embeddings ensures generalization across 23 languages, outperforming CNN-only or MFCC-based baselines.
- **Practical Deployment Feasibility:** With lightweight preprocessing and offline evaluation capability, the model is suitable for real-world multilingual audio deepfake detection systems.

#### IV. RESULTS AND DISCUSSIONS

The proposed XLS-R + Conformer hybrid model was evaluated on the MLAAD dataset, a multilingual benchmark containing offline audio samples across 23 languages. The dataset includes diverse manipulation types such as Text-to-Speech (TTS), Voice Conversion (VC), and splicing/editing, recorded under clean, noisy, and compressed conditions. The dataset was split into training, validation, and test sets, with equal numbers of real and fake audio clips to ensure unbiased evaluation. The results demonstrate that the hybrid model achieves superior detection accuracy, robustness, and cross-lingual generalization compared to conventional CNN, RNN, and MFCC-based methods, while efficiently handling offline audio distortions.

##### A. Performance Comparison

The performance of the proposed model was compared against popular baseline architectures, including CNN-LSTM and ResNet-Transformer.

*Table 1: Performance comparison of different models on multilingual offline audio deepfake detection*

Model	Accuracy	Precision	Recall
CNN-LSTM	94.73%	91.5%	91.8%
ResNet-Transformer	92.1%	94.2%	91.9%
XLS-R + Conformer (Ours)	98.45%	98.5%	98.0%

##### Observations:

The proposed model outperforms CNN-LSTM and ResNet-Transformer architectures across all metrics.

Improvements are attributed to the combination of multilingual embeddings from XLS-R, temporal modeling from Conformer, and feature fusion, which captures both local micro-patterns (glitches, unnatural harmonics) and global fluency patterns (prosody, rhythm).

Traditional CNN-only or RNN-only models lack the capability to detect subtle cross-lingual speech manipulations, which explains their lower accuracy.

##### B. Confusion Matrix and Class-Wise Analysis

A detailed class-wise performance analysis provides insight into the model's reliability:

###### Class 0 (Real Audio, Bonafide):

- Precision: 98% (Test), 99% (Validation)
- Recall: 98% (Test), 99% (Validation)
- F1-Score: 98% (Test), 99% (Validation)

###### Class 1 (Fake Audio, Spoof):

- Precision: 99% (Test), 98% (Validation)
- Recall: 98% (Test & Validation)
- F1-Score: 98% (Test & Validation)

##### Observations:

The confusion matrix shows minimal misclassifications, indicating high reliability in differentiating between genuine and fake audio.

Balanced performance across both classes ensures that the system is robust against false positives (real audio misclassified as fake) and false negatives (fake audio misclassified as real).

The model achieves high recall, minimizing missed detections of manipulated audio—a critical factor in security-sensitive applications.

##### C. Macro and Weighted Averages

- Macro Average: Precision: 98.5%, Recall: 98% (Test), F1-Score: 98%
- Weighted Average: Precision: 98.4%, Recall: 98.4% (Test), F1-Score: 98.4%

These metrics highlight the model's consistent performance across both classes, accounting for potential class imbalances in the dataset. The weighted averages further demonstrate robustness in real-world scenarios, where the distribution of genuine and fake audio may not be equal.

#### D. Insights and Observations

- **Cross-Lingual Generalization:** XLS-R embeddings provide pretrained multilingual knowledge, enabling the model to detect deepfakes across 23 languages without retraining. This is a significant advantage over monolingual CNN or RNN models.
- **Robust Offline Detection:** Preprocessing, including noise reduction, silence trimming, and augmentation, ensures that the model is effective in offline playback conditions with background noise, compression, or device-specific artifacts.
- **Balanced Class Performance:** The F1-scores indicate a well-balanced trade-off between precision and recall for both real and fake audio. The model minimizes both false negatives and false positives, critical for trust in security or forensic applications.
- **Detection of Subtle Manipulations:** By combining XLS-R embeddings with Conformer's temporal attention, the model identifies subtle synthesis artifacts such as unnatural pauses, pitch drifts, and inconsistencies in prosody or rhythm.
- **Advantages Over Baselines:**
  - CNN-LSTM: Limited temporal modeling reduces performance on long-duration and offline audio.
  - ResNet-Transformer: Focused primarily on spatial features, missing subtle temporal cues.
  - XLS-R + Conformer: Multi-level feature fusion captures both global multilingual patterns and local temporal inconsistencies, yielding the highest accuracy and F1-scores.

#### E. Comparative Evaluation

- Baseline models (CNN-LSTM and ResNet-Transformer) achieve accuracies between 88%–95% on offline audio deepfake datasets.
- XLS-R + Conformer achieves 98.45% test accuracy, showing substantial improvement in real/fake discrimination.
- The hybrid architecture enables robust detection of all manipulation types (TTS, VC, splicing) in noisy and compressed offline conditions, which traditional baselines struggle to handle.

#### F. Deployment and Practical Considerations

- The model's lightweight preprocessing and offline evaluation make it suitable for deployment on CPU-based devices, such as laptops and embedded systems.
- Efficient computation ensures low latency during offline processing, enabling practical applications in digital forensics, fraud detection, and offline voice authentication systems.
- The model's cross-lingual design reduces retraining costs for new languages or domains.

#### G. Future Directions

- **Enhancing Fake Audio Recall:** Explore ensemble learning or adversarial training to improve detection of more sophisticated or previously unseen spoofing attacks.
- **Dataset Expansion:** Adding more offline audio samples across additional languages and dialects will strengthen the model's generalization capabilities.
- **Real-Time Adaptation:** Although currently designed for offline evaluation, optimization for ultra-low latency may enable near-real-time applications while maintaining cross-lingual accuracy.
- **Explainability and Forensics:** Integrating interpretability techniques can allow investigators to identify specific manipulated segments in offline audio samples.

## V. CONCLUSION

This paper proposes a hybrid XLS-R + Conformer framework for offline multilingual audio deepfake detection. The system leverages cross-lingual self-supervised embeddings extracted by XLS-R to capture detailed phonetic, prosodic, and speaker-specific features across 23 languages. These embeddings are combined with the Conformer network, which models both local acoustic variations and long-term temporal dependencies, enabling the detection of subtle and sophisticated manipulations in speech.

The preprocessing pipeline, including noise reduction, resampling to 16 kHz, silence trimming, amplitude normalization, and data augmentation, ensures that the model effectively handles offline audio recorded under diverse conditions, including clean, noisy, and compressed environments. By

fusing global and temporal cues through the XLS-R and Conformer modules, the system achieves high robustness against a variety of deepfake generation techniques, such as Text-to-Speech (TTS), Voice Conversion (VC), and splicing/editing attacks.

Experimental results on the MLAAD dataset demonstrate superior performance, with a test accuracy of 98.45% and validation accuracy of 98.91%, alongside high precision, recall, and F1-scores for both real and fake audio classes. The model consistently maintains balanced performance across classes, showcasing effective generalization across multiple languages and recording conditions. These results highlight the effectiveness of combining multilingual embeddings and temporal modeling in a unified framework, providing a reliable, offline solution for real-world deepfake audio detection.

In summary, the proposed hybrid XLS-R + Conformer approach not only outperforms conventional architectures such as CNN-LSTM and ResNet-Transformer but also provides practical applicability, making it suitable for offline deployment in security, digital forensics, and voice authentication applications.

## VI. FUTURE WORK

Despite achieving near-optimal performance, several improvements and research directions can further enhance the system's capabilities:

- **Expansion to Low-Resource Languages and Dialects:** While the current model supports 23 languages, extending it to include low-resource languages and regional dialects can enhance cross-lingual generalization, enabling the detection of deepfakes in previously unsupported linguistic contexts.
- **Enhanced Fake Audio Detection via Ensemble and Adversarial Training:** Incorporating ensemble learning or adversarially trained models can strengthen the detection of sophisticated fake audio, improving recall for subtle manipulations and reducing false negatives.
- **Real-Time Offline Deployment Optimization:** The model can be further optimized for ultra-low-latency execution on edge devices, such as smartphones or embedded IoT devices, ensuring practical integration into voice authentication

systems, forensic tools, and offline monitoring applications.

- **Multimodal Deepfake Detection Integration:** Future work may involve combining audio detection with video, lip movement, or facial synchronization analysis to create a multimodal deepfake detection system, improving robustness against complex audiovisual manipulations.
- **Explainable AI and Visualization of Detection Cues:** Incorporating explainability techniques can provide end-users with insights into why certain audio samples are flagged as fake, increasing trust and transparency in forensic and security applications.
- **Dynamic Dataset Expansion and Continuous Learning:** Continuous integration of newly generated deepfake datasets and retraining or fine-tuning the model can improve its adaptability to emerging spoofing techniques, maintaining high accuracy over time.

By exploring these directions, the XLS-R + Conformer framework can evolve into a highly reliable, practical, and deployable system for multilingual offline audio deepfake detection, capable of addressing both current and future challenges in digital audio security.

## REFERENCES

- [1] E. Jain and A. Singh, "Deepfake voice detection using convolutional neural networks: A comprehensive approach to identifying synthetic audio," in 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS), 2024.
- [2] Z. Jin, L. Lang, and B. Leng, "Wave-Spectrogram Cross-Modal Aggregation for Audio Deepfake Detection," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, doi: 10.1109/ICASSP49660.2025.10890563.
- [3] R. Mahyavanshi, C. V. Mahesh Reddy, A. J. Shah, and H. A. Patil, "Teager Energy Cepstral Coefficients for Audio Deepfake Detection," in 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2024, doi: 10.1109/APSIPAASC63619.2025.10848893.

- [4] P. Chiddarwar, "Real-Time Detection of AI-Generated Deepfake Audio: A Novel Approach," in 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG), 2024, doi: 10.1109/ICTBIG64922.2024.10911062.
- [5] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio Deepfake Approaches," IEEE Access, vol. 11, 2023, doi:10.1109/ACCESS.2023.3333866.
- [6] M. Gujjar, A. U. Rehman, K. Munir, M. Amjad, and A. Bermak, "Unmasking the Fake: Machine Learning Approach for Deepfake Voice Detection," IEEE Access, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3521026.
- [7] M. Gaikawad and S. Ghosh, "A robust and lightweight CNN-Transformer model for audio deepfake detection in Indian languages," in 2025 7th International Conference on Signal Processing, Computing and Control (ISPCC), 2025. DOI: 10.1109/ISPCC66872.2025.11039572.
- [8] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "BTS-E: Audio deepfake detection using breathing-talking-silence encoder," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023. DOI: 10.1109/ICASSP49357.2023.10095927.
- [9] G. S. Kashyap, S. Kumar, Z. H. Siddiqui, N. Kamuni, M. A. Azeez, J. Gao, and R. Ali, "Fooling the forgers: A multi-stage framework for audio deepfake detection," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025. DOI: 10.1109/ICASSP49660.2025.10888175.
- [10] D. J. Dsouza, A. P. Rodrigues, and R. Fernandes, "Multi-Modal Comparative Analysis on Audio Dub Detection Using Artificial Intelligence," IEEE Access, vol 18, 2025, doi: 10.1109/ACCESS.2025.3591306.
- [11] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "A efficient temporary deepfake location approach based embeddings for partially spoofed audio detection," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024. DOI: 10.1109/ICASSP48485.2024.10448196.
- [12] D. Song, N. Lee, J. Kim, and E. Choi, "Anomaly detection of deepfake audio based on real audio using generative adversarial network model," IEEE Access, 2024.
- [13] D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth, and S. Tubaro, "Audio Splicing Detection and Localization Based on Acquisition Device Traces," IEEE Transactions on Information Forensics and Security, vol. 18, 2023.