

Research paper

Multilingual code-switched spoofed speech detection using self-supervised learning models

Hafsa Ilyas^a, Muhammad Hamza^b, Junaid Mir^{b,*,*}, Ali Javed^a, Muhammad Haroon Yousaf^c, Ahmed Zoha^d, Shah Nawaz^e, Habib Ullah Manzoor^d

^a Department of Software Engineering, University of Engineering & Technology Taxila, Taxila, Pakistan

^b Department of Electrical Engineering, University of Engineering & Technology Taxila, Taxila, Pakistan

^c Department of Computer Engineering, University of Engineering & Technology Taxila, Taxila, Pakistan

^d James Watt School of Engineering, University of Glasgow, Glasgow, United Kingdom

^e Institute of Computational Perception, Johannes Kepler University Linz, Linz, Austria

ARTICLE INFO

Keywords:

Code-switching

Audio deepfake detection

Spoofed speech dataset

Application of artificial intelligence

Implemented artificial intelligence

ABSTRACT

The rapid advancement of audio generation technologies has enabled the creation of highly realistic synthetic speech, underscoring the need for robust audio spoof detection techniques. However, most existing countermeasures are evaluated on monolingual datasets with utterances from native speakers, leaving their performance on code-switched multilingual speech underexplored. To address this gap, we introduce the multilingual spoofed speech (MSS) dataset, featuring bona fide and spoofed utterances in Urdu, Hindi, and English that capture the natural conversational and code-switching patterns of South Asian speakers. A fused self-supervised learning (SSL) embedding-based audio anti-spoofing framework using an integrated spectro-temporal graph attention network (AASIST) is proposed that combines representations extracted from waveform-to-vector 2.0 and universal speech representation with speaker-aware training frontends. Extensive experiments are conducted on the proposed MSS dataset and the benchmark automatic speaker verification spoofing and countermeasures (ASVspoof) 2019 logical access (LA), ASVspoof-2021 LA, and ASVspoof-2021 deepfake (DF) datasets. Cross-lingual and cross-corpora evaluations demonstrate that the proposed anti-spoofing framework achieves excellent generalization, attaining equal error rates of 0.38%, 0.99%, and 2.46% on ASVspoof-2019 LA, ASVspoof-2021 LA, and ASVspoof-2021 DF, respectively. Moreover, multilingual training on MSS significantly improves the robustness of the implemented artificial intelligence (AI) model, particularly in handling code-switched and linguistically diverse audio. Overall, the results highlight both the engineering contribution of the MSS dataset and the AI contribution of the proposed multi-SSL AASIST framework, together advancing the application of AI for cross-lingual generalizability and the reliable detection of spoofed speech across monolingual and code-switched scenarios.

1. Introduction

The advancement of artificial intelligence (AI) has revolutionized audio spoofing attacks to such an extent that spoofed audio can now deceive both humans and automatic speaker verification (ASV) systems. Audio spoofing attacks involve creating realistic fake audio using AI-based algorithms and are generally categorized into text-to-speech (TTS) and voice conversion (VC) attacks. TTS synthesis generates speech from text with natural prosody and articulation patterns, whereas VC attacks manipulate the original speech to mimic a target speaker's voice while preserving the linguistic content. In addition to AI-driven spoofing methods, another notable category is replay

attacks, in which a voice biometric system is deceived using pre-recorded samples of the target speaker. The cumulative progress in audio spoofing technologies has created a serious social threat, as spoofed audio can be exploited to spread misinformation, propaganda, and fraud. For instance, an individual's spoofed voice can bypass ASV systems, amplifying security risks as voice-based utilities become increasingly widespread. Moreover, spoofed voices can facilitate impersonation attempts, scams, social engineering, financial fraud, and online harassment. Consequently, detecting spoofed audio has gained significant attention in the research community as an essential measure to mitigate these attacks and enhance societal security.

* Corresponding author.

E-mail address: junaid.mir@uettaxila.edu.pk (J. Mir).

<https://doi.org/10.1016/j.engappai.2026.116398>

Received 28 October 2025; Received in revised form 16 September 2026; Accepted 24 September 2026

Available online 3 October 2026

0952-1976/© 2026 Elsevier Ltd. All rights reserved, including those for text and data mining, AI training, and similar technologies.

To mitigate the security threats faced by ASV systems, the research community introduced the ASVspoof Challenge (ASVspoof Challenge, 2025), inviting researchers to develop effective countermeasures (CMs) against voice spoofing. Since 2015, the ASVspoof Challenge has released datasets biannually, focusing primarily on TTS, VC, and replay attacks. However, these datasets are based on English-language audio samples. Beyond the ASVspoof Challenge, the research community has contributed several other datasets for audio spoofing detection. For instance, WaveFake (Frank and Schönherr, 2021) includes TTS-generated samples in English and Japanese, while TIMIT-TTS (Salvi et al., 2023) comprises synthetic speech utterances produced using twelve TTS algorithms. Similarly, the CFAD dataset (Ma et al., 2024) contains synthetic audio samples in Chinese, generated through twelve different TTS systems. These datasets have substantially facilitated research on spoofing CMs. Despite this progress, most existing spoofing datasets consist of audio samples from native English speakers and fail to capture the English-speaking styles of South Asians, particularly Urdu and Hindi speakers. Moreover, these datasets include only monolingual audio utterances (i.e., each utterance contains only one language) and overlook the code-switching tendencies of Urdu–English and Hindi–English bilingual speakers, who frequently switch between English and Urdu or Hindi, respectively, during conversation. Therefore, there is a clear need for a spoofing dataset that accurately reflects the conversational style and English-speaking accent of bilingual individuals, especially Urdu and Hindi speakers, who collectively represent a population of over half a billion.

Over the past few years, numerous machine learning (ML) and deep learning (DL) approaches have been developed for the ASVspoof challenges, achieving significant performance improvements and state-of-the-art (SOTA) results. However, most SOTA approaches struggle to generalize to out-of-domain scenarios, particularly in cross-corpora settings. Recently, self-supervised learning (SSL) models for speech representation, such as HuBERT (Hsu et al., 2021), Wav2Vec2 (Baeovski et al., 2020a), and WavLM (Tran et al., 2024), have attracted significant attention in audio deepfake detection. Trained on large-scale raw and unlabeled speech data, these feature extraction models can extract richer and more discriminative speech representations than traditional handcrafted features such as mel-frequency cepstral coefficients (MFCCs) and linear-frequency cepstral coefficients (LFCCs) (Tran et al., 2024). Among DL architectures designed for spoofing detection, the AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal graph attention networks) framework (weon Jung et al., 2022) has shown competitive performance due to its ability to model complex spectro-temporal dependencies in speech signals. Despite these advancements, challenges remain in developing SSL-based spoofing detection systems that generalize well across diverse datasets and languages. Furthermore, the potential of multi-model SSL-based frontends that extract fused feature representations has not yet been thoroughly explored in this domain.

In this paper, we introduce a multilingual spoofed speech (MSS) dataset to address the aforementioned limitations of existing audio spoofing datasets. The proposed MSS dataset comprises code-switched audio utterances from Urdu and Hindi speakers in Urdu, Hindi, and English, thereby representing the natural conversational and English-speaking style of South Asians. In addition to this dataset, we propose AASIST-based frameworks incorporating SSL-based feature extractors. Specifically, we introduce SSL frontends into the AASIST architecture to analyze the impact of fused features from multiple SSL models on spoofed utterance detection. The investigated SSL model combinations include Wav2Vec2+UniSpeech-SAT, WavLM+UniSpeech-SAT, and Wav2Vec2+WavLM. Among these, the Wav2Vec2+UniSpeech-SAT-based AASIST framework demonstrates superior performance. The main contributions of the paper are as follows:

- Proposed a multilingual spoofed speech (MSS) dataset that includes code-switched multilingual audio samples capturing diverse speaking styles and mimicking the natural conversational dynamics of Urdu–English and Hindi–English bilingual speakers.

- Developed a multi-embedding AASIST-based spoof detection framework incorporating different SSL-based feature extractors to analyze the impact of fused features on spoofed audio detection.
- Conducted extensive experiments, including cross-lingual and cross-corpora evaluations, using the proposed MSS dataset alongside standard benchmarks such as the ASVspoof-2019 LA, ASVspoof-2021 LA, and ASVspoof-2021 DF datasets.

The rest of the paper is structured as follows: the literature review is presented in Section 2, and the proposed methodology is presented in Section 3. Details of the datasets used, evaluation measures, and experimental settings are given in Section 4, with results along with a detailed discussion presented in Section 5. Finally, the conclusion is provided in Section 6.

2. Literature review

This section presents a review of existing methods for audio spoofing detection and discusses the available audio spoofing datasets.

2.1. Audio spoofing detection approaches

Several audio spoofing detection approaches have been introduced in the literature. Initially, the research community focused on conventional feature-engineering-based approaches for audio spoofing detection. Later, researchers explored end-to-end DL models that process raw audio. Spectrogram-based approaches are also highlighted in the literature; however, the most recent research in the audio spoofing detection domain focuses on SSL-based methods. The subsequent subsections discuss conventional and DL-based audio spoofing detection approaches.

2.1.1. Conventional machine learning approaches

Early research on audio spoofing detection relied on conventional ML approaches that primarily used handcrafted feature extractors and ML classifiers to mitigate the threats posed by spoofing attacks in ASV systems. Acoustic ternary patterns, combined with gammatone cepstral coefficients, were introduced in Javed et al. (2021) to capture the distortion signature and unnatural prosodic artifacts in replay and cloned samples. A support vector machine (SVM) classifier was used for classification. This approach was evaluated on VSDC and ASV-19 logical access (LA) and physical access (PA) datasets and reported an equal error rate (EER) of 0.1 on the ASV-19 LA dataset. Extending this work, a unified framework based on an acoustic ternary co-occurrence pattern feature extractor was presented in Javed et al. (2022). This work employed error-correcting output codes and combined three binary SVM classifiers to classify the extracted features. An extended local ternary pattern (ELTP) feature descriptor, combined with LFCC, was used in Arif et al. (2021) to detect LA attacks. The ELTP-LFCC features were classified using a deep BiLSTM, achieving an EER of 0.74 on the ASV-19 LA dataset. The spoofing detection approach based on central lopsided local binary patterns with an LSTM classifier was demonstrated in Dawood et al. (2022), with a reported EER of 0.06 on the ASV-19 LA dataset. A study in Hamza et al. (2022) presented an approach that classified MFCC features through an SVM classifier and evaluated it on the FakeorReal dataset. In Rahmeni et al. (2022), acoustic and glottal flow descriptors were presented to extract pertinent features, which were then classified using SVM and XGBoost classifiers for voice spoofing detection. Traditional ML methods are computationally less intensive and achieve satisfactory performance on individual datasets. However, the performance of such approaches has not been assessed in cross-corpora settings to validate their generalization ability.

2.1.2. Deep learning-based audio spoofing detection methods

DL-based approaches have been extensively explored to address the limitations of conventional ML methods in audio spoofing detection. These methods can broadly be categorized into three groups: (i) models that process raw audio waveforms, (ii) models that operate on time-frequency representations such as spectrograms, and (iii) SSL-based frameworks.

Approaches processing raw waveforms. In the audio spoofing detection domain, RawNet (weon Jung et al., 2019) is a prominent convolutional neural network (CNN) architecture that directly processes raw waveforms. In RawNet, the convolutional layer is applied directly to the raw audio signal; however, higher layers in the architecture include residual blocks with skip connections and a gated recurrent unit layer to aggregate utterance-level representations. RawNet2 (Tak et al., 2021a) is an advanced version of RawNet, featuring and incorporating a Sinc convolutional layer that implements parameterized Sinc functions to process raw audio directly and incorporates feature map scaling in the residual block as an attention mechanism. RawNet2 reported a minimum tandem detection cost function (min-tDCF) of 0.129 and an EER of 4.66. Based on RawNet2, the research community has introduced several other spoofing CMs. For instance, the RawNet2-based model StateNet (Ranjan et al., 2022) learns spatiotemporal features to detect spoofing attacks. StateNet was evaluated on ASV-19 LA, in-the-wild, and FoR-Norm datasets. The framework introduced in Sun et al. (2023) implemented the vocoder identification module in RawNet2 and assessed its performance on ASV-19 LA, WaveFake, and their own LibriSeVoc dataset. The performance of these methods (Ranjan et al., 2022; Sun et al., 2023) on the standard ASV-21 LA/DF dataset, including the attack-wise evaluation, is unknown. A meta-learning framework incorporating RawNet2 was demonstrated in Wang and Hansen (2024) to improve generalization. The model was evaluated on the ASV-19 LA dataset and achieved a min-tDCF and an EER of 0.0277 and 0.87, respectively. Attack-wise evaluation was also performed to demonstrate generalizability to unseen spoofing attacks. However, cross-corpus evaluation is needed to further validate the generalization ability.

To capture both spectral and temporal representations from the waveform and improve the spoofing detection performance, graph attention networks (GATs) have been utilized. The work in Tak et al. (2021c) demonstrated that the GAT-based solutions outperform the baseline RawNet2 for detecting spoofing attacks. Building on the work in Tak et al. (2021c), RawGAT-ST (Tak et al., 2021b) was presented, integrating a spectral-temporal GAT with RawNet2 to fuse spectral and temporal representations. A min-tDCF of 0.0335 and an EER of 1.06 were reported on the ASV-19 LA dataset along with the attack-wise evaluation. To further improve performance, the AASIST (weon Jung et al., 2022) framework was introduced, which implements heterogeneous stacking graph attention layers and readout mechanisms. For ASV-19 LA, AASIST attained a min-tDCF and an EER of 0.0275 and 0.83, respectively. The generalizability of the above-described GAT-based approaches (Tak et al., 2021a; weon Jung et al., 2022) is uncertain, as these methods have not been evaluated in cross-corpora settings, on ASV-21 LA/DF, or on other spoofing datasets.

Spectrogram image-based approaches. The effectiveness of spectrogram image-based methods for detecting spoofing attacks has also been highlighted in the literature. VGG16 and WaveletCNN were used in Fathan et al. (2022) with an additive-margin softmax loss to classify audio spectrogram images. However, the reported performance on the ASV-19 LA and DF datasets requires further improvement. In Wang et al. (2023), a joint training approach was introduced that incorporates a speech enhancement module (employing a UNet) alongside LCCN- and ResNet-based audio spoofing detection models. The method processed audio spectrograms, was evaluated on the ASV-19 LA and FAD datasets, and demonstrated robustness in noisy scenarios with low signal-to-noise ratios. AVFakeNet, introduced in Ilyas et al. (2023), was a Swin

Transformer-based architecture that used mel-spectrogram images and reported an EER of 0.13 on the ASV-19 LA dataset. However, performance on the ASV-21 dataset is unknown, leaving uncertainty about generalization performance. Likewise, a method in Goel et al. (2024) presented a vision transformer-based approach and reported an EER of 4.74 on the ASV-21 LA dataset. The generalizability of the methods (Wang et al., 2023; Goel et al., 2024) is uncertain, as is their attack-wise performance. To enhance generalization, a study in Le et al. (2024) introduced the audio spectrogram transformer, comprising 12 transformer encoder blocks that processed audio spectrogram images. The model was trained on training data gathered from various speech generation sources and then evaluated on the ASV-19 LA, in-the-wild, and FakeAVCeleb datasets. Based on audio spectrograms, Febrinanto et al. (2025) presented a spatio-temporal vision graph non-contrastive learning (SIGNL) framework that employs vision graph convolutional encoders pre-trained with a non-contrastive learning strategy. The SIGNL framework attained EERs of 7.21, 2.33, and 8.44 on the ASV-21 DF, ASVspoof-5, and CFAD datasets, respectively. Cross-corpora evaluation was also conducted using an in-the-wild dataset; however, attack-wise evaluation is needed to assess the approach's effectiveness, particularly for TTS and VC-based attacks.

SSL-based approaches. Nowadays, the research paradigm for audio spoofing detection has shifted toward SSL-based frontends to improve the generalization of spoofing CMs. To demonstrate the importance of SSL models for audio spoofing detection, two SSL-based frontends (Wav2Vec2 and HuBERT) were investigated in Wang and Yamagishi (2021) in combination with three distinct backend classifiers (a classifier dubbed LLGF and its two variants). The Wav2Vec2-XLS-R with LLGF showed the best performance with EERs of 2.31, 7.18, 5.44, and 0.24 on the ASV-19 LA, ASV-21 LA, ASV-21 DF, and ASV-5 datasets, respectively. The Wav2Vec feature extractor with a classifier that includes a temporal normalization and statistical pooling layer was introduced in Martín-Doñas and Álvarez (2022). The reported EERs for ASV-21 LA and ASV-21 DF were 3.54 and 4.98, respectively. In Tak et al. (2022), a spoofing CM using Wav2Vec2 with AASIST was demonstrated. For the ASV-21 LA dataset, the model achieved an EER and a min-tDCF of 0.82 and 0.2099; however, for ASV-21 DF, an EER of 2.85 was reported. The attack-wise performance of the methods (Martín-Doñas and Álvarez, 2022; Tak et al., 2022) needs to be investigated for the ASV-21 LA dataset, along with evaluation on other spoofing datasets. In Pan et al. (2024), the WavLM architecture was integrated with LSTM and ECAPA-TDNN models. The embeddings from the early layers of WavLM with the LSTM model achieved the best EERs of 0.65, 3.50, and 3.19 on ASV-19 LA, ASV-21 LA, and ASV-21 DF datasets, respectively. Likewise, a study in Kheir et al. (2025) conducted a layer-wise analysis of SSL models, namely Wav2Vec2, WavLM, and HuBERT, along with two backend classifiers, a feedforward neural network and AASIST. These models were evaluated across multiple datasets and languages (English, Chinese, and Spanish) and on partially spoofed utterances. The studies (Pan et al., 2024; Kheir et al., 2025) highlight that lower layers of SSL models yield better results than using the entire model. AASIST2 was presented in Zhang et al. (2024), in which the Sinc layer was modified with Wav2Vec2, the ResNet blocks were replaced with Res2Net, and additive angular margin softmax was incorporated. EERs of 0.15, 1.61, and 2.77 were reported on the ASV-19 LA, ASV-21 LA, and ASV-21 DF datasets.

The method in Sinha et al. (2024) investigated the fusion of embeddings from SSL models (HuBERT and Wav2Vec2) within a supervised learning framework for spoofing detection. This approach was evaluated on ASV-19 LA and ASV-21 LA and attained min-tDCF values of 0.0053 and 0.2504, respectively. However, the results for the ASV-21 DF dataset need to be evaluated to effectively validate the approach's generalization ability (Sinha et al., 2024). An interpretable framework implementing Wav2Vec2-based frame-level and attentive utterance-level features was introduced in Li and Zhang (2024). The model was

trained on three classes (TTS, VC, and bona fide), and temporal class activation maps were used for interpretability. A min-tDCF and an EER of 0.0165 and 0.51, respectively, were reported for the ASV-19 LA dataset. However, this interpretable model was not evaluated on the ASV-21 LA and DF datasets, leaving its generalizability uncertain. A lightweight nested Res2Net backend over a Wav2Vec2 frontend (Nes2Net) was proposed in Liu et al. (2025) with only 511 K backend parameters, reporting EERs of 1.61 and 1.89 on ASV-21 LA and DF, respectively. RawNetLite (Di Pierno et al., 2025), an end-to-end raw-waveform model, was trained using domain mixing, focal loss, and waveform-level augmentation, achieving an in-domain EER of 0.0029 on FakeOrReal. XLS-R and BEATs representations were fused within AASIST in Zaman et al. (2026), obtaining an EER of 2.99 on Comp-SpoofV2. Non-semantic TRILL/TRILLsson representations with LSTM, multi-head attention, and MLP back-ends were explored in Das et al. (2025), reporting an EER of 0.96 on ASV-19 LA but showing degraded in-the-wild performance (EER of 27.52). CA2Net (Sun et al., 2025) combined coordinate attention with Res2Net, achieving EERs of 1.42 and 0.67 on the ASV-19 LA and PA sets, respectively. TRACE (Farooq et al., 2026) was a training-free CM that analyzed embedding trajectories from a frozen WavLM model for partial-spoof detection. RawNet-based CMs such as DeepLSD (Al-Tairi et al., 2025) and multi-head-attention-refined Wav2Vec-AASIST variants (Viakhirev et al., 2025) have also been investigated on ASVspoof and ASVspoof-5 data. A recent study proposed the JMDS corpus (Mawalim et al., 2026) (17 languages, over 350,000 utterances) and benchmarked several detectors, with an XLS-R-AASIST model achieving the best EER of 8.22, followed by WavLM-AASIST at 10.97. XMAD-Bench (Ciobanu et al., 2026) provided a cross-domain multilingual benchmark across seven languages and multilingual dataset integration strategies for the SAFE Challenge (Ali et al., 2025). The proposed model improved the in-the-wild EER to 8.46 using a WavLM trained on five datasets.

While recent SOTA SSL-based audio spoofing detection models achieve excellent performance on benchmark datasets, cross-dataset evaluations consistently report performance degradation. Moreover, existing work has primarily focused on English and a few other high-resource languages, with little or no evaluation on Urdu and Hindi. Given the distinct phonetic, prosodic, and code-switching characteristics of these languages, the generalizability of current SSL-based spoof detection models remains largely unexplored, underscoring the need for systematic evaluation and the development of more robust multilingual spoof detection approaches.

2.2. Audio spoofing datasets

To accelerate research on spoofing CMs, researchers have introduced several spoofing-detection datasets. Among them, the ASVspoof datasets released biannually under the ASVspoof Challenge are prominent and are considered standard in the research community. ASVspoof-2015 (Wu et al., 2015) was the first edition of the challenge, whereas the latest edition is ASVspoof-5 (Wang et al., 2024). ASVspoof-2015 incorporated VC and speech synthesis (SS) attacks, and the second edition, ASVspoof-2017 (ASV-17) (Kinnunen et al., 2017), focused on replay attacks (RA). However, ASVspoof-2019 (ASV-19) (Wang et al., 2020) introduced two categories: logical access (LA) attacks and physical access (PA) attacks. The LA track included TTS and VC attacks, whereas the PA track included replay attacks. ASVspoof-2021 (ASV-21) (Yamagishi et al., 2021) introduced a third category, deepfake (DF) attacks, in addition to LA and PA attacks. The latest edition, ASVspoof-5, also included adversarial attacks alongside TTS and VC attacks. The audio deep synthesis detection (ADD) challenge (Yi et al., 2022) was introduced in 2022 and included three tracks: low-quality fake audio detection (LF), partially fake audio detection (PF), and audio fake game (FG). The ASV challenge focused on the English language, whereas the ADD challenge was based on Mandarin

Chinese. In addition, the literature encompasses many other audio-spoofing datasets that advance research on audio spoofing detection. For example, the FakeorReal (FoR) dataset (Reimao and Tzerpos, 2019) included bona fide and spoofed samples generated using diverse TTS systems. The study in Müller et al. (2022) introduced an in-the-wild dataset comprising audio samples of English-speaking politicians and celebrities. This dataset (Müller et al., 2022) included spoofed and bona fide utterances (with background noise) gathered from social platforms to reflect real-world spoofing scenarios. The FMFCC-a dataset (Zhang et al., 2021) included spoofed Chinese audio generated using different VC and TTS algorithms. The HABLA dataset (Flórez et al., 2023) incorporated TTS- and VC-generated spoof samples in Spanish. The MLAAD dataset (Müller et al., 2024) covered 23 languages (excluding Urdu) and provided only spoofed samples generated using diverse TTS approaches. An overview of the existing spoofing detection datasets is provided in Table 1.

The literature on audio spoofing datasets indicates that most existing corpora primarily feature English speakers and focus on English. A few datasets include Chinese-language samples, while only a limited number support multiple languages. Moreover, most datasets generate spoofed samples primarily via TTS algorithms, with little to no inclusion of VC samples. Notably, no publicly available dataset captures the speaking accents and linguistic patterns of South Asian speakers, particularly Urdu- and Hindi-speaking individuals. Furthermore, existing datasets overlook the code-switching behavior commonly observed in bilingual speakers who alternate between English and their native languages during conversation. These gaps underscore the need for a dedicated audio spoofing dataset that reflects South Asian English-speaking accents and code-switching behavior, thereby providing more realistic, representative conditions for developing robust spoofing detection systems.

3. Methodology

In this section, we introduce the proposed AASIST framework, which uses multiple SSL frontends rather than a Sinc convolutional layer to effectively detect audio deepfakes. SSL models have attracted considerable attention for their ability to capture comprehensive speech characteristics and learn meaningful, generalized representations from raw audio. Fig. 1 provides an overview of the proposed approach for audio spoof detection. The raw input audio is fed to a multi-SSL feature extractor, followed by a post-processing block, a residual encoder, a self-attention aggregation layer, and a spectro-temporal graph attention network. The extracted representations are finally classified as either bona fide or spoofed speech. The description of each module is provided in the subsections that follow.

3.1. SSL feature extractors

To extract SSL representations from raw audio, multiple SSL-based frontends are integrated and investigated within the AASIST framework. The proposed framework leverages features from multiple SSL models, which are fused via channel-wise concatenation of embeddings. This fusion exploits the complementary strengths of different SSL models, enhancing the framework's ability to capture comprehensive, diverse, and generalized speech representations. Mathematically, the SSL feature fusion can be expressed as:

$$\text{ssl}_{f_x} = \text{SSL}_{M_x}(I), \quad (1)$$

$$\text{ssl}_{f_y} = \text{SSL}_{M_y}(I), \quad (2)$$

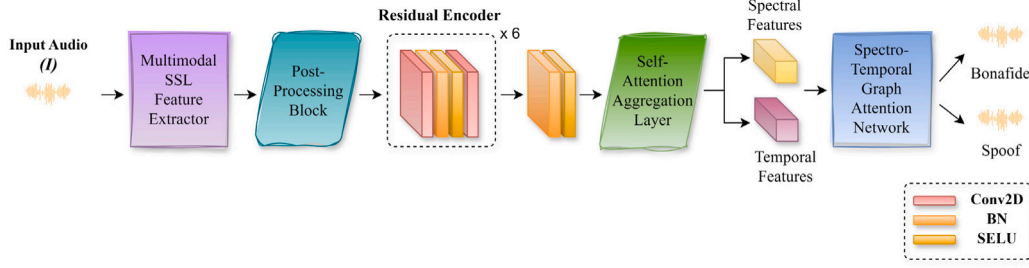
$$\text{ssl}_{f_m} = \text{ssl}_{f_x} \parallel \text{ssl}_{f_y}, \quad (3)$$

where the input audio signal is represented as I , ssl_{f_x} refers to the features extracted through the first SSL model denoted by SSL_{M_x} .

Table 1

Overview of existing audio spoofing and deepfake detection datasets.

Dataset	Year	Language	Spoofed methods	# Real	# Fake	Male	Fem.
ASVspoof-2015 (Wu et al., 2015)	2015	English	TTS = 3, VC = 7	16,651	246,500	45	61
ASVspoof-2017 (Kinnunen et al., 2017)	2017	English	–	3566	14,466	42	
ASVspoof-2019 (Wang et al., 2020)	2019	English	LA: TTS = 13, VC = 6 PA	12,483 218,430	109,816	46	61
FakeorReal (Reimao and Tzerpos, 2019)	2019	English	TTS = 7	111,000	87,285	33	
ASVspoof-2021 (Yamagishi et al., 2021)	2021	English	LA: TTS = 13, VC = 6 DF PA	14,816 14,869 94,068	133,360 519,059 627,264	30 43 30	37 50 37
FMFCC-a (Zhang et al., 2021)	2021	Mandarin Chinese	TTS = 11, VC = 2	10,000	40,000	58	73
WaveFake (Frank and Schönherr, 2021)	2021	English, Japanese	TTS = 7	0	117,985	0	2
ADD 2022 (Yi et al., 2022)	2022	Mandarin Chinese	–	493,123		128	
In-the-wild (Müller et al., 2022)	2022	English	–	31,779		58	
HABLA (Flórez et al., 2023)	2022	Spanish	TTS = 3, VC = 3	22,816	58,000	78	84
TIMIT-TTS (Salvi et al., 2023)	2023	English	TTS = 12	0	5160	46	
CFAD (Ma et al., 2024)	2024	Chinese	TTS = 12	38,600	77,200	1212	
MLAAD (Müller et al., 2024)	2024	38	TTS = 82	0	154,000	—	
ASVspoof-5 (Wang et al., 2024)	2024	English	TTS = 19, VC = 6, AT = 7	171,901	815,262	964	958

**Fig. 1.** Overview of the proposed multiple SSL model-based audio spoofing detection architecture.

Similarly, the features extracted using the second SSL model SSL_{M_y} are represented as ssl_{f_y} . The fused feature vector is denoted as ssl_{f_m} .

We investigate combinations of Wav2Vec2 (Baevski et al., 2020b), UniSpeech-SAT (Chen et al., 2022a) and WavLM (Chen et al., 2022b) SSL models. The Wav2Vec2 (Baevski et al., 2020b) model is composed of a multi-layer CNN feature encoder and a transformer-based context network that contextualizes the latent representations to learn speech patterns. UniSpeech-SAT (Chen et al., 2022a) is also a transformer-based SSL model, optimized for speaker-aware learning by integrating utterance-wise contrastive learning. It is a multitask learning framework that incorporates connectionist temporal classification (CTC) and contrastive loss. WavLM (Chen et al., 2022b) consists of a CNN encoder and a transformer network with gated relative position bias to capture speech sequence ordering. It further incorporates masked speech denoising, making it robust against noise and complex acoustic environments. The Wav2Vec2 model can capture linguistic, phonetic, and temporal features, which are crucial for detecting spoofing artifacts, such as unnatural transitions and vocoder artifacts. Meanwhile, the UniSpeech-SAT model extracts speaker-specific features alongside linguistic and contextual representations. In contrast, WavLM captures spectral, phonetic, prosodic, and noise-robust representations, making it particularly suitable for noisy environments. As illustrated in Fig. 2, different SSL model combinations are investigated, including Wav2Vec2+WavLM, Wav2Vec2+UniSpeech-SAT, and WavLM+UniSpeech-SAT. The fused feature vector derived from these combinations encompasses comprehensive representations that can enhance the framework's performance in detecting spoofed speech. The feature map ssl_{f_m} is processed by a post-processing block comprising a layer that adds a channel dimension, max pooling, batch normalization (BN), and a SeLU activation

layer. The post-processing block transforms the fused feature map into spectro-temporal representations R_m , which are subsequently fed to the residual encoder.

3.2. Residual encoder and self-attention aggregation layer

A residual encoder comprising six residual blocks is employed, where each residual block consists of a 2D convolutional (Conv) layer followed by BN, SeLU activation function, and another 2D Conv layer. The residual encoder enables the network to extract high-level representations $F_m \in \mathbb{R}^{c \times s \times t}$, where the number of channels, spectral bins, and time-frequency frames are represented as c , s , and t , respectively. The operation of the residual encoder can be expressed as:

$$F_m = \text{Conv2D}(\text{BN}(\text{SeLU}(\text{Conv2D}(R_m)))) \quad (4)$$

A self-attention-based aggregation layer is then implemented to extract relevant spectral and temporal representations. A 2D weighted attention matrix AM_w is generated by processing the representations F_m through a Conv layer followed by SeLU, BN, another Conv layer, and a Softmax function that normalizes the weights. The weighted sum of representations F_m is then computed across the time (t) and spectral (s) domains as:

$$T_F = \sum_s (F_m \odot AM_w), \quad (5)$$

$$S_F = \sum_t (F_m \odot AM_w), \quad (6)$$

where, T_F and S_F refer to the time and spectral representations, respectively, while $\odot$ denotes element-wise multiplication. The resulting representations are finally passed to the spectro-temporal graph attention network for further processing.

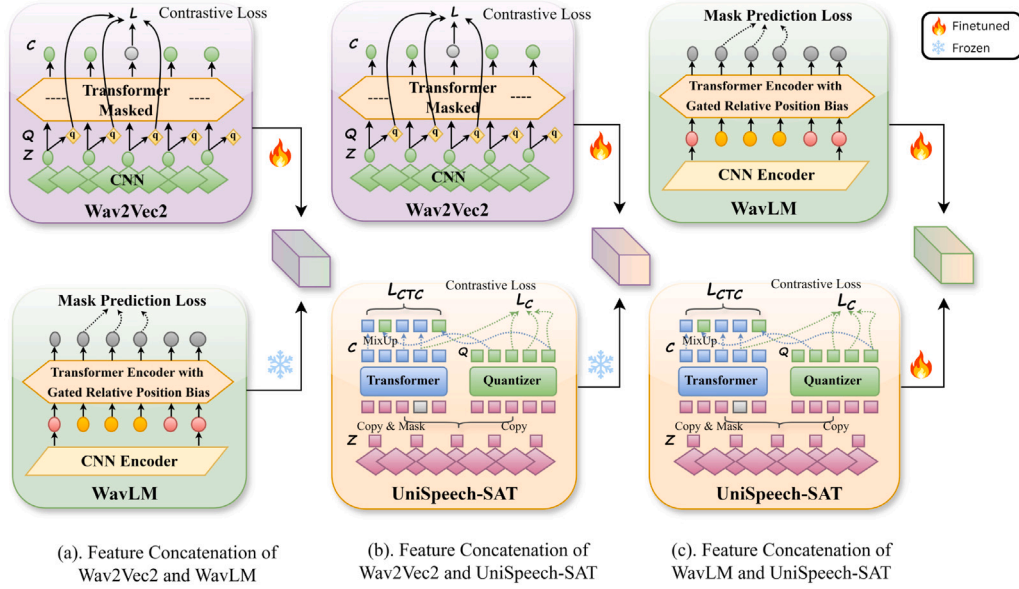


Fig. 2. Multimodal SSL feature extractor configurations evaluated in this study. Each configuration combines two SSL models by concatenating their extracted embeddings to exploit complementary speech representations. The fused feature representations are subsequently forwarded to the downstream spoof detection network for classification. The three investigated combinations are (a) Wav2Vec2 with WavLM, (b) Wav2Vec2 with UniSpeech-SAT, and (c) WavLM with UniSpeech-SAT.

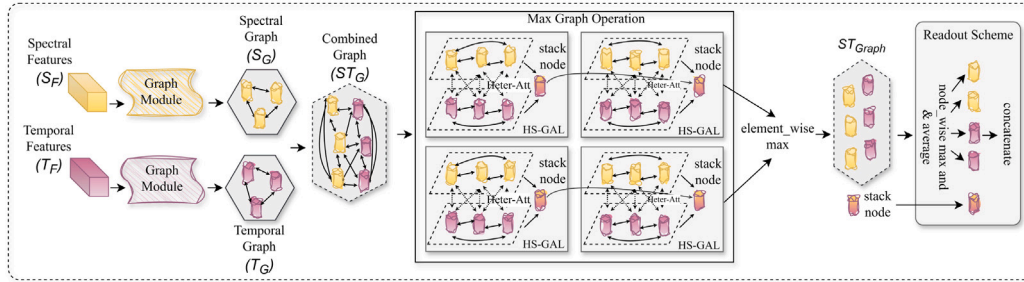


Fig. 3. The architecture of the spectro-temporal graph attention network, illustrating how the spectral (S_F) and temporal (T_F) representations are converted into graphs, combined into a heterogeneous spectro-temporal graph, processed by the max graph operation (MGO) with heterogeneous stacking graph attention layers, and aggregated by the readout scheme.

3.3. Spectro-temporal graph attention network

The spectro-temporal graph attention network (illustrated in Fig. 3) includes a pair of graph modules for the spectral S_F and temporal T_F representations, max graph operation (MGO), and a readout scheme. Each graph module comprises the graph attention network and graph pooling (GP) layer. The obtained spectral S_F and temporal T_F representations are processed through corresponding graph modules to construct spectral S_G and temporal T_G graphs, containing the set of graph nodes. Mathematically, spectral S_G and temporal T_G graphs are computed as:

$$S_G = \text{GraphModule}(\max_s(|S_F|)), \quad (7)$$

$$T_G = \text{GraphModule}(\max_t(|T_F|)). \quad (8)$$

The S_G and T_G are then combined to construct a heterogeneous spectro-temporal graph ST_G , having the nodes of both spectral and temporal graphs but in different latent spaces. The heterogeneous ST_G is then processed by the MGO. The MGO consists of two branches, each comprising two heterogeneous stacking graph attention layers (HS-GALs) followed by two GP layers. HS-GAL includes the heterogeneous attention mechanism to accommodate graph heterogeneity and stack

node to accumulate information between spectral s and temporal t domains. HS-GAL in each branch of MGO shares the common stack node to preserve the information in S_G and T_G . After that, the element-wise maximum operation is applied to the outputs from the parallel branches of MGO to generate another heterogeneous spectro-temporal graph ST_{Graph} . This ST_{Graph} is processed through a readout scheme, where the readout layer extracts four nodes via applying node-wise maximum and average operations to the nodes belonging to S_G and T_G . These four nodes are then concatenated with the stack node, whereas the final prediction (bona fide and spoofed) is generated through the last hidden fully connected layer.

4. Experiments

This section provides details on the description of datasets, evaluation measures, and experimental settings.

4.1. Datasets

To evaluate the performance, we considered the standard ASVspoof-2019 LA (Wang et al., 2020), ASVspoof-2021 LA/DF (Yamagishi et al., 2021), and our proposed multilingual spoofed speech (MSS) datasets. A brief description of each dataset is provided in the following subsections.

Table 2

Statistics of the MSS dataset showing the speaker-disjoint splits for the Urdu, Hindi, and English language subsets, including the number of bona fide and spoofed samples, duration in hours (h), and male/female speaker distribution.

Dataset	Split	Bona fide #Samples (h)	Spoofed #Samples (h)	Male	Female
MSS – Urdu	Train	3418 (15.7)	15,300 (67.6)	11	7
	Dev	3857 (16.1)	13,600 (53.2)	11	6
	Eval	5431 (22.4)	59,500 (239.2)	21	14
MSS – Hindi	Train	2287 (9.7)	20,500 (90.9)	14	7
	Dev	2520 (10.4)	21,000 (83.4)	14	7
	Eval	3329 (13.1)	85,895 (361.1)	28	14
MSS – English	Train	5126 (20.1)	36,300 (147.0)	25	14
	Dev	5625 (23.4)	34,520 (137.5)	25	13
	Eval	8678 (36.8)	145,600 (637.4)	49	28
Total	–	40,271 (167.7)	432,215 (1817.3)		

Table 3

Complexity and computational-cost comparison of the evaluated AASIST-based spoof detection frameworks. Inference time is reported for a 15-s input utterance.

Configuration	Total parameters	Trainable parameters	Time per epoch	Time 25 epochs	Inference
Wav2Vec2 – AASIST (baseline)	315.9 M	315.9 M	≈30 min	≈12.5 h	≈90 ms
WavLM + UniSpeech-SAT – AASIST	189.3 M	189.3 M	≈20 min	≈8.5 h	≈60 ms
Wav2Vec2 + WavLM – AASIST	410.4 M	316.0 M	≈35 min	≈14.6 h	≈150 ms
Wav2Vec2 + UniSpeech-SAT – AASIST	410.3 M	316.0 M	≈40 min	≈16.7 h	≈140 ms

4.1.1. Multilingual spoofed speech (MSS) dataset (proposed)

The MSS dataset is a multilingual speech collection with bona fide and spoofed utterances in Urdu, Hindi, and English. Bona fide samples are collected from the MAV-Celeb dataset (Nawaz et al., 2021), which features celebrities' bona fide audio only from YouTube videos and captures real-world conditions such as background noise, language switching, and multi-speaker environments. The MSS dataset includes 154 bilingual speakers (70 Urdu and 84 Hindi). The Urdu and Hindi subsets contain utterances in their respective native languages, while the English subset contains English utterances spoken by the same Urdu and Hindi speakers. Table 2 presents the statistics of the proposed MSS dataset. The dataset is created using a speaker-disjoint partitioning protocol with a 25:25:50 split for the training, validation, and test sets, respectively. Specifically, the split was performed at the speaker level while preserving the original gender distribution. All utterances belonging to a given speaker were assigned to the same partition, ensuring that each speaker appeared in only one split. Consequently, there is no speaker overlap between the training, validation, and test sets. Overall, the MSS dataset comprises approximately 1985 h of audio (approximately 167.7 h bona fide and 1817.3 h spoofed).

Spoofed samples in the MSS dataset are generated using five VC models, namely FreeVC (Li et al., 2023), Diff-HierVC (Choi et al., 2023), HierSpeech++ (Lee et al., 2023), kNN-VC (Baas et al., 2023), and SeedVC (Liu, 2024). VC algorithms are specifically used to preserve the conversational dynamics of South Asian Urdu and Hindi speakers while generating spoofed audio in the target speakers' voices. Spoofed samples are generated independently within each speaker-disjoint split to prevent any cross-split speaker leakage. For a split containing N speakers, 10 utterances were randomly selected from each speaker as source utterances. For each source speaker, one target utterance was randomly selected from each of the remaining $N - 1$ speakers. The 10 source utterances were then individually converted into the target speaker's voice using the selected target utterance as the reference. This process was repeated for every source–target speaker pair and across all five VC techniques, yielding $N \times (N - 1) \times 10 \times 5$ spoofed samples per split. The per-split totals in Table 2 follow this procedure, with minor differences where a few source utterances were unavailable or individual conversions failed.

The proposed MSS dataset captures the conversational dynamics of bilingual individuals who switch between English and their native language. The MSS dataset specifically includes audio samples with

code-switching that reflect the conversational style of South Asians (Urdu and Hindi speakers). Additionally, the English set of the MSS dataset covers the English-speaking style of Urdu and Hindi speakers, along with variations in speaking style and pronunciation across individuals. The existing audio spoofing datasets lack these characteristics, underscoring the uniqueness and importance of the proposed MSS dataset. Overall, the MSS dataset includes audio samples with code-switching, multilingual conversational patterns, and diversity in English-speaking style, accent, and pronunciation. The MSS dataset is publicly available on Kaggle as three language subsets: MSS–English,¹ MSS–Urdu,² and MSS–Hindi.³ Each language subset is partitioned into speaker-disjoint training, development, and evaluation splits and includes the corresponding protocol files. The subsets contain both bona fide and spoofed audio samples in FLAC format with an average length of 15 s and a 16 kHz sampling rate.

4.1.2. ASVspoof-2019 LA dataset

The ASV-19 LA (Wang et al., 2020) comprises the spoofed and bona fide samples of 107 speakers, split into training, development, and evaluation sets. There are 2580 bona fide and 22,800 spoofed utterances in the training set, and 1064 bona fide and 22,296 spoofed samples in the development set. The evaluation set contains 7355 bona fide and 63,882 spoofed utterances. Specifically, the training and development sets contain 6 known attacks (A01–A06) generated using 4 TTS and 2 VC techniques, whereas the evaluation set comprises 2 known and 11 unknown attacks generated using 7 TTS and 6 VC algorithms.

4.1.3. ASVspoof-2021 LA and DF dataset

The ASV-21 LA and DF datasets (Yamagishi et al., 2021) comprise only evaluation sets containing bona fide and spoofed utterances, with no new training or development sets. ASV-21 DF includes 14,869 bona fide and 519,059 spoofed samples from 93 speakers, generated using over 100 spoofing algorithms. This dataset comprises compressed deep-fake speech processed with various lossy codecs used for media storage, thus simulating social media content. In particular, spoofed audio is

¹ <https://www.kaggle.com/datasets/mssdataset/mss-english>

² <https://www.kaggle.com/datasets/mssdataset/mss-urdu>

³ <https://www.kaggle.com/datasets/mssdataset/mss-hindi>

encoded under nine codec conditions, including dual-codec scenarios, to enhance realism. ASV-21 LA dataset consists of 121,156/14,816 bona fide and 638,982/133,360 spoofed utterances across the progress and evaluation subsets, with spoofed samples generated using 13 TTS, VC, and hybrid attack algorithms. The ASV-21 LA dataset is more challenging than ASV-19 LA because it includes telephony encoding and transmission artifacts across seven codec conditions, thereby introducing compression, packet loss, and bandwidth variations.

4.2. Evaluation measures

To evaluate performance, we used the standard metrics for audio spoofing detection: equal error rate (EER) and tandem detection cost function (min-tDCF) (Wang et al., 2020; Yamagishi et al., 2021). The EER represents the point where the false acceptance rate (FAR) becomes equal to the false rejection rate (FRR). It reflects the trade-off between accepting spoofed speech and rejecting bona fide utterances; therefore, a lower EER indicates better model performance. The min-tDCF measures the overall system cost based on the false rejections and false acceptances by ASV and spoofing CM systems, and incorporates prior probabilities and error costs for target speakers, impostors, and spoofing attacks. It represents the combined performance of ASV and CM systems.

4.3. Experimental settings

The proposed model is implemented using the Python PyTorch framework. It was trained on two classes (spoofed and bona fide) with a batch size of 8 and 25 epochs. The other model training hyperparameters are a weighted categorical cross-entropy loss, the Adam optimizer, a 1×10^{-6} learning rate, and a 1×10^{-4} weight decay. The model checkpoint is saved at each epoch, and the checkpoint with the lowest validation loss is selected for evaluation. Specifically, we employed the Wav2Vec2 XLS-R 300M multilingual checkpoint (xlr2_300m; ≈ 315.4 M parameters, 1024-dimensional embeddings, 24 transformer layers), the UniSpeech-SAT base checkpoint (unispeech-sat-base; ≈ 94.4 M parameters, 768-dimensional embeddings, 12 transformer layers), and the WavLM base checkpoint (wavlm-base; ≈ 94.4 M parameters, 768-dimensional embeddings). A fixed random seed of 1234 was used in all experiments to ensure reproducibility and a controlled comparison across configurations. All experiments are conducted on an RTX 3090 GPU machine.

4.4. Computational complexity

Table 3 reports a comparison of the complexity and computational cost between the evaluated AASIST-based dual-SSL configurations and the baseline. In configurations involving Wav2Vec2 (i.e., Wav2Vec2+UniSpeech-SAT and Wav2Vec2+WavLM), only Wav2Vec2 was fine-tuned while the second SSL encoder was kept frozen. This design choice allows Wav2Vec2 to adapt its multilingual speech representation to spoof-discriminative characteristics while preserving the complementary pretrained representation of the second SSL encoder. Moreover, keeping the secondary encoder frozen maintains a number of trainable parameters (≈ 316.0 M) comparable to that of the baseline Wav2Vec2-AASIST model (≈ 315.9 M), enabling a more controlled comparison of the benefit of incorporating complementary SSL representations. However, while incorporating a second SSL encoder increases inference latency to ≈ 140 – 150 ms compared to the baseline (≈ 90 ms), reflecting the additional forward-pass computation, the observed latency remains relatively low for practical audio spoof detection applications. In the WavLM+UniSpeech-SAT configuration, both WavLM and UniSpeech-SAT were fine-tuned, as this configuration has substantially fewer total parameters (≈ 189.3 M) than the Wav2Vec2-based dual-SSL configurations and the baseline.

5. Results and discussion

The performance of the proposed AASIST frameworks incorporating multiple SSL feature extractors is evaluated through two experiments. In the first experiment, the models are assessed on the proposed MSS dataset, where training and testing are conducted separately on each language subset (MSS – Urdu, MSS – Hindi, and MSS – English). Specifically, for each language subset, the training and development sets are used for training and validation, respectively, while the evaluation set is used for testing. In the second experiment, the standard ASVspoof-2019 LA (ASV-19 LA) dataset is utilized to analyze the performance of different combinations of SSL feature extractors. For this experiment, the models are trained using the training and development sets of the ASV-19 LA dataset and evaluated on the evaluation set. The results of both experiments are provided in Table 4.

The results show that all models exhibit higher EER on the MSS – Hindi, indicating the challenging nature of Hindi utterances. The elevated EER reflects the models' difficulty in distinguishing bona fide from spoofed samples within the Hindi language subset of the proposed MSS dataset. This may be attributed to frequent code-switching in the Hindi set, which involves phonetic and prosodic shifts, as well as overlapping voices and background music, which introduce noise into the spoofed samples. Models achieve better results on the ASV-19 LA dataset than on the MSS – English, highlighting the greater complexity of MSS due to its code-switched utterances and intra-sample language alternation. In contrast, the monolingual nature and silent segments of ASV-19 LA may contribute to its comparatively lower EER.

Among the evaluated models, Wav2Vec2+UniSpeech-SAT demonstrates consistently strong performance across MSS sets and the ASV-19 LA dataset, achieving the best results on the MSS – Hindi and MSS – English, and a relatively low min-tDCF of 0.0094 on ASV-19 LA. Wav2Vec2+WavLM performs best on the Urdu subset and ASV-19 LA, whereas WavLM+UniSpeech-SAT shows the poorest performance, with notably higher EERs. Overall, the findings underscore the robustness of the Wav2Vec2+UniSpeech-SAT framework in distinguishing code-switched spoofed and bona fide utterances across multiple languages. Consequently, this framework is selected for further experimentation.

5.1. Attack-wise evaluation

We conducted an attack-wise evaluation to analyze the performance of the selected framework (Wav2Vec2+UniSpeech-SAT) against individual spoofing methods, thereby assessing its resilience across different attack types. This evaluation was performed on both the standard ASV-19 LA and the proposed MSS dataset in a two-stage experiment.

In the first stage, an attack-wise analysis was conducted on the MSS dataset, following the same training protocols. The trained model for each language subset (Urdu, Hindi, and English) was tested separately on individual attacks, and the resulting EER values are reported in Table 5. The results indicate that the SeedVC attack poses the greatest challenge, yielding the highest EER values across all MSS language subsets. This suggests that SeedVC generates spoofed samples with characteristics that are particularly difficult for the framework to detect. The HierSpeech++ attack is also notably challenging, achieving the second-highest EERs of 5.22 and 1.4 for the Hindi and English subsets, respectively. In contrast, FreeVC and Diff-HierVC generated samples are the easiest for the framework to identify, with comparatively lower EER across all languages. These findings demonstrate that the framework's effectiveness varies across spoofing algorithms and languages. Overall, the proposed Wav2Vec2+UniSpeech-SAT framework achieves the best performance on the MSS – Urdu, consistently yielding lower EER across all attacks. On the other hand, the MSS – Hindi remains the most challenging, with the highest EER across all spoofing methods.

In the second stage, we evaluated the attack-wise performance of the proposed Wav2Vec2+UniSpeech-SAT framework on the ASV-19 LA

Table 4

Performance comparison of different SSL model combinations on the MSS (Urdu, Hindi, English) and ASVspoof-2019 LA datasets.

Models	MSS			ASV-19 LA	
	Urdu (EER (%))	Hindi (EER (%))	English (EER (%))	min-tDCF	EER (%)
WavLM + UniSpeech-SAT	11.51	15.50	6.14	0.2924	3.78
Wav2Vec2 + WavLM	0.25	8.26	2.61	0.0079	0.32
Wav2Vec2 + UniSpeech-SAT	0.43	5.55	1.39	0.0094	0.38

Table 5

Attack-wise EER (%) of the proposed Wav2Vec2+UniSpeech-SAT framework on the MSS dataset.

Attacks	MSS - Urdu	MSS - Hindi	MSS - English
FreeVC	0.07	5.05	1.40
kNN-VC	0.48	4.45	1.41
Diff-HierVC	0.13	5.03	0.68
HierSpeech++	0.20	5.22	1.42
SeedVC	0.99	7.39	1.71

Table 6

Attack-wise evaluation of the proposed Wav2Vec2+UniSpeech-SAT framework on the ASVspoof-2019 LA dataset.

Attacks	min-tDCF	EER (%)
A07	0.00098	0.04
A08	0.00909	0.37
A09	0.00000	0.00
A10	0.00791	0.29
A11	0.00452	0.15
A12	0.00237	0.08
A13	0.00000	0.00
A14	0.00081	0.04
A15	0.00981	0.04
A16	0.00118	0.04
A17	0.00705	0.34
A18	0.02633	0.77
A19	0.01153	0.49

Table 7

Cross-lingual evaluation of the proposed Wav2Vec2+UniSpeech-SAT framework.

Train set	Test set	EER (%)
MSS – Urdu	MSS – Hindi	10.25
	MSS – English	5.01
	MSS – Urdu	0.43
MSS – Hindi	MSS – Urdu	3.43
	MSS – English	8.02
	MSS – Hindi	5.55
MSS – English	MSS – Urdu	2.44
	MSS – Hindi	5.47
	MSS – English	1.39
ASV-19 LA (English)	MSS – Urdu	19.74
	MSS – Hindi	30.16
	ASV-19 LA	0.38

dataset. The model was trained once using the complete training and development sets of the ASV-19 LA dataset and subsequently evaluated separately for each spoofing attack on the held-out evaluation set. The results obtained in terms of min-tDCF and EER are summarized in Table 6. The evaluation subset of the ASV-19 LA dataset includes TTS attacks (A07–A16) and VC attacks (A17–A19), with all attacks except A16 and A19 being unknown (unseen during training). Among these, A17 is considered particularly challenging, as prior spoofing CM studies have reported relatively poor performance on this attack. The results reveal that the proposed framework exhibits excellent performance in detecting TTS-based spoofing, achieving particularly strong results on the A09 and A13 attacks. Furthermore, it demonstrates robust detection capabilities for VC-based spoofing, attaining the lowest min-tDCF (0.00705) and EER (0.34%) for the most challenging A17 attack. These results emphasize the effectiveness and generalizability of the proposed Wav2Vec2+UniSpeech-SAT framework in detecting diverse spoofing techniques, including previously unseen attacks.

5.2. Cross-lingual evaluation

To investigate the cross-lingual generalizability and language dependency of the proposed Wav2Vec2+UniSpeech-SAT framework, we conducted an experiment where the model trained on one language subset (e.g., Urdu) of the MSS dataset was evaluated on the other subsets (Hindi and English), and vice versa. The results of these experiments are presented in Table 7. The findings reveal a noticeable decline in performance in cross-language evaluations compared to in-language evaluations, underscoring the model's language-specific learning behavior. Specifically, the model trained on the Urdu subset attained the highest EER of 10.25 when tested on the Hindi subset, indicating limited cross-lingual generalization from Urdu to Hindi. Conversely, the model trained on the Hindi subset achieved the lowest EER of 3.43 when evaluated on the Urdu subset. This improvement can be attributed to the partial lexical overlap between Hindi and Urdu. While Urdu primarily derives from Persian and Arabic, Hindi originates

Table 8

Cross-corpora evaluation of the proposed Wav2Vec2+UniSpeech-SAT framework.

Train dataset	Test dataset	min-tDCF	EER (%)
ASV-19 LA	ASV-21 LA	0.2124	0.99
	ASV-21 DF	–	2.46
	MSS – English	–	23.93
MSS – English	ASV-21 LA	0.5093	15.21
	ASV-21 DF	–	12.78
	ASV-19 LA	0.3805	9.52

Table 9

Performance comparison of different fusion mechanisms in the proposed Wav2Vec2+UniSpeech-SAT framework on the MSS dataset.

Fusion mechanisms	MSS		
	Urdu (EER (%))	Hindi (EER(%))	English (EER (%))
Weighted fusion	1.06	7.63	1.81
Gated Fusion	1.23	9.99	2.57
Cross-attention Fusion	2.25	35.89	2.91
Concatenation (Proposed)	0.43	5.55	1.39

from Sanskrit and incorporates several Persian and Arabic loanwords. This overlap facilitates some shared learning of phonetic and acoustic representations, thereby enhancing the model's cross-lingual transfer performance from Hindi to Urdu. Furthermore, the model trained on the MSS – English subset demonstrates strong cross-lingual adaptability, achieving EERs of 2.44 and 5.47 for the Urdu and Hindi subsets, respectively. This superior generalization may be attributed to the code-switched nature of the MSS dataset, in which the English subset includes utterances that alternate between English–Urdu and English–Hindi segments. Such code-switching exposes the model to diverse phonetic, acoustic, and linguistic patterns, thereby improving its ability to detect spoofed audio across multiple languages and mixed-language utterances.

An additional experiment was also conducted in which the proposed Wav2Vec2+UniSpeech-SAT framework was trained on the ASV-19 LA dataset and evaluated on the Urdu and Hindi subsets of the MSS dataset, as presented in the last row of Table 7. The model achieved a high EER of 19.74 on the MSS–Urdu and 30.16 on the MSS–Hindi. This significant degradation in performance can be attributed to the phonetic and acoustic dissimilarities between English (the language of ASV-19 LA) and the target languages (Urdu and Hindi). The elevated EER values also reflect the challenges posed by cross-dataset variations, including code-switching patterns, linguistic diversity, speaker accent, and pronunciation differences between datasets. Overall, the findings indicate that the model trained on code-switched audio samples (as in the MSS dataset) exhibits superior cross-lingual generalization, whereas a model trained on a monolingual English dataset struggles to generalize to multilingual, code-switched speech. These results emphasize the language dependency of spoofing CMs and highlight the complexity of detecting audio deepfakes in code-switched, multilingual scenarios where prosodic variations, phonetic nuances, and articulation patterns differ substantially across languages.

5.3. Cross-corpora evaluation

To conduct a cross-corpora evaluation of the proposed Wav2Vec2+UniSpeech-SAT framework, we used the ASV-19 LA, ASV-21 LA/DF, and MSS – English datasets. First, the model was trained on ASV-19 LA and evaluated on ASV-21 LA, ASV-21 DF, and the MSS–English set. Second, the model was trained on the MSS–English set and assessed on the ASV-21 LA, ASV-21 DF, and ASV-19 LA datasets. The results of both experiments are summarized in Table 8.

The model trained on ASV-19 LA achieved excellent EERs of 0.99 and 2.46 when evaluated on ASV-21 LA and ASV-21 DF, respectively. However, it attained the highest EER of 23.93 on the MSS–English set. These results indicate that the model generalizes well to detect

codec transmission variations in ASV-21 LA and compressed spoofed audios in ASV-21 DF, yet struggles to generalize across datasets with distinct linguistic and phonetic characteristics. This limitation arises because ASV-19 LA is monolingual, representing native English speech, whereas the MSS – English set consists of code-switched audio samples from Urdu- and Hindi-speaking speakers, with English and their native language co-occurring in a single utterance. For a model trained on the MSS – English set, EERs of 15.21 and 12.78 are observed on the ASV-21 LA and ASV-21 DF datasets, respectively. This performance degradation can be attributed to the MSS dataset not incorporating codec transmissions or compression artifacts present in the ASVspoof datasets. Interestingly, the model performed better on ASV-21 DF than on ASV-21 LA, demonstrating its robustness against unseen compressed deepfake samples. When evaluated on ASV-19 LA, the model attained a satisfactory EER of 9.52 under cross-corpora conditions.

The performance drop is noticeable across entirely distinct datasets such as ASV and MSS, mainly due to differences in spoofing techniques and domain mismatches, including linguistic variability, transmission channels, and codec compression. ASV datasets contain monolingual English speech, whereas the MSS dataset incorporates bilingual code-switched utterances. Such linguistic domain mismatch involves prosodic, phonetic, and lexical variations, resulting in shifts in acoustic feature distributions that can reduce the transferability of learned decision boundaries. Moreover, the ASVspoof datasets notably include TTS-based attacks unseen by the model trained on the MSS dataset, which includes only VC-based attacks. The underlying reason lies in the fundamental difference between TTS and VC attacks. TTS synthesizes speech directly from text, often introducing unnatural prosody and robotic intonations that make it easier to detect. In contrast, VC transforms existing human speech while preserving natural prosody and acoustic dynamics, modifying only the speaker's identity. This retention of naturalness makes VC-generated deepfakes substantially more challenging to detect due to their subtle and less systematic artifacts.

Overall, the results suggest that, in the evaluated cross-corpora setting, the model trained on code-switched audio generalizes better to the monolingual datasets, despite the absence of TTS attacks and codec-compressed speech during training. As indicated, the model trained on the MSS – English set attains an EER of 9.52 on ASV-19 LA, compared to an EER of 23.93 attained by the model trained on ASV-19 LA and evaluated on the MSS – English set. This finding suggests the potential benefit of diverse, linguistically rich training data in enhancing model robustness. Overall, the proposed Wav2Vec2+UniSpeech-SAT framework demonstrates satisfactory performance across diverse datasets in cross-corpora settings, indicating its ability to learn transferable feature representations.

5.4. Ablation study

This section presents ablation studies designed to examine the contribution of different fusion strategies to the overall performance of the proposed framework, analyze the impact of compressed audio samples on detection performance, and assess the effect of multilingual training with a linguistically diverse dataset. The details are provided in the subsequent subsections.

5.4.1. Impact of fusion mechanisms

This ablation study investigates the effectiveness of the fusion strategy adopted in the proposed framework for combining Wav2Vec2 and UniSpeech-SAT representations. In particular, we compare simple concatenation with three advanced fusion mechanisms, including weighted fusion, gated fusion, and cross-attention. Let ssl_{f_x} and ssl_{f_y} denote the representations from Wav2Vec2 and UniSpeech-SAT, respectively. For simple concatenation, the representations are combined along the feature dimension and can be mathematically expressed as:

$$ssl_{f_m} = ssl_{f_x} \parallel ssl_{f_y}, \quad (9)$$

where $\parallel$ denotes the concatenation operation.

In the weighted fusion mechanism, separate learnable parameters are introduced for the Wav2Vec2 and UniSpeech-SAT channels, and a sigmoid transformation is applied to the weights before feature scaling. The learnable coefficients are used to modulate the Wav2Vec2 and UniSpeech-SAT representations prior to concatenation. The weighted fusion is computed as:

$$ssl_{f_w} = \gamma_{f_x} ssl_{f_x} \parallel \gamma_{f_y} ssl_{f_y}, \quad (10)$$

where γ_{f_x} and γ_{f_y} denote the learnable feature-wise weights.

In gated fusion, a gating vector is learned from the joint SSL representations and is used to regulate Wav2Vec2 features. The gate is generated using a non-linear network consisting of fully connected layers, ReLU activation, and a sigmoid. The gated fusion mechanism can be mathematically represented as:

$$ssl_{f_g} = g \odot ssl_{f_x} \parallel ssl_{f_y}, \quad (11)$$

where g indicates the gating vector and $\odot$ denotes element-wise multiplication.

The cross-attention fusion explicitly models the dependencies between the two SSL representations. Specifically, the Wav2Vec2 embedding is used as the query, while the UniSpeech-SAT embedding serves as the key and value. Both representations are projected into a common 768-dimensional latent space, followed by four-head multi-head attention. Mathematically represented as:

$$ssl_{f_c} = CA(ssl_{f_x}, ssl_{f_y}) \parallel ssl_{f_y}, \quad (12)$$

where $CA(\cdot)$ represents the cross-attention operation.

Each model configuration was evaluated under the same experimental protocols using the MSS dataset. Table 9 summarizes the results of the conducted ablation study in terms of EER. The results demonstrate that the simple concatenation (proposed framework) achieved the lowest EER among all the fusion mechanisms across each language subset of the MSS dataset. Notably, the cross-attention fusion results in substantially higher EERs, specifically for the MSS – Hindi subset. The observed performance degradation of all other fusions may be associated with the additional transformations they impose on SSL representations. Weighted fusion modifies feature contribution through learned scaling, potentially reducing the influence of discriminative information. Gated fusion dynamically controls feature propagation, which can suppress complementary spoofing cues. Cross attention emphasizes features based on learned cross-dependencies between the two SSL representations, which may not capture spoofing cues that are independently informative. Moreover, these sophisticated fusion mechanisms introduce additional parameters, thereby increasing computational complexity. On the other hand, simple concatenation avoids

these limitations by retaining complete information from both SSL models without additional parameters. This indicates that sophisticated fusion mechanisms may offer greater representation flexibility; however, they do not improve audio spoofing detection performance in the evaluated setting, despite their increased complexity. The ablation study reveals that simple concatenation yields the most effective fusion of complementary SSL representations among the evaluated strategies within the proposed framework for the audio spoofing detection task.

5.4.2. Impact of compression

To analyze the effect of compressed audio samples on spoofing detection performance, we conducted experiments under three conditions using the Urdu, Hindi, and English language subsets of the MSS dataset. For each language subset, the proposed Wav2Vec2+UniSpeech-SAT framework, trained on the corresponding training and development subsets, was evaluated on the evaluation set under all three conditions. In the first experiment, the evaluation utterances were used in their original form with no compression. In the second experiment, different compression levels were applied to the bona fide and spoofed audio samples in the evaluation set, whereas in the third experiment, compression was applied only to the spoofed utterances. The compression included MP3 at 96 and 240 kbps, advanced audio coding (AAC) at 24 and 96 kbps, OGG at quality levels q3 and q8, and two sequential compression settings: MP3 followed by AAC and OGG followed by AAC. The results are presented in Table 10 and reveal that compression has a substantial effect on spoofing detection performance. It can be observed that when compression is applied to both spoofed and bona fide utterances, the EERs increase to 0.90, 9.64, and 2.10 for the Urdu, Hindi, and English subsets, respectively. In this setting, the compression artifacts are shared across spoofed and bona fide classes, potentially altering the acoustic distribution of the evaluation data relative to that seen during training. Consequently, this distribution shift may reduce the effectiveness of the learned discriminative representations, resulting in higher EER. Notably, when compression is applied only to spoofed utterances, spoofing detection performance improves for the Urdu and English subsets; however, it remains unchanged for the Hindi subset of the MSS dataset. This suggests that applying compression only to spoofed utterances may introduce additional discriminative cues, increasing class separability and resulting in lower EERs. Overall, this ablation study demonstrates that the spoofing detection framework is sensitive to compression-induced acoustic variations, with the impact varying across languages and compression scenarios.

5.4.3. Impact of multilingual training data

To analyze the impact of multilingual training using a larger and linguistically diverse dataset, we conducted an experiment in which the proposed Wav2Vec2+UniSpeech-SAT framework was trained on the entire MSS dataset. For this purpose, the training and development sets of each language subset (Urdu, Hindi, and English) were combined to train the model. The trained framework was then evaluated on each language subset and on different combinations of language subsets within the MSS dataset. The results reported in Table 11 indicate a clear improvement in model robustness across all individual subsets, suggesting that multilingual training enhances the framework's ability to learn rich cross-linguistic representations and generalize effectively across Urdu, Hindi, and English code-switched utterances. Notably, the EER reduction on the Hindi set from 5.55 to 2.99 demonstrates the benefit of multilingual training for the more challenging MSS–Hindi subset, which contains more frequent code-switching and background music interference. Furthermore, the framework achieved EERs of 1.18 and 0.86 when evaluated on the combined English+Urdu and English+Hindi sets, respectively. When tested on a combined set of audio samples from Urdu, Hindi, and English, the model achieved the lowest EER of 0.76, underscoring the importance of a linguistically diverse, code-switched dataset for improving the performance and generalization of audio spoofing detection frameworks.

Table 10

Impact of compression on the performance of the proposed Wav2Vec2+UniSpeech-SAT framework.

	MSS		
	Urdu (EER (%))	Hindi (EER(%))	English (EER (%))
w/o Compression	0.43	5.55	1.39
w/ Compression	0.90	9.64	2.10
w/ Compression (only Spoof)	0.33	5.55	1.21

Table 11

Evaluation of the proposed Wav2Vec2+UniSpeech-SAT framework trained on the complete MSS dataset on individual and combinations of MSS language subsets.

Dataset(s)	EER (%)
MSS – Urdu	0.37
MSS – Hindi	2.99
MSS – English	0.81
MSS – English + MSS – Urdu	1.18
MSS – English + MSS – Hindi	0.86
MSS – English + MSS – Urdu + MSS – Hindi	0.76

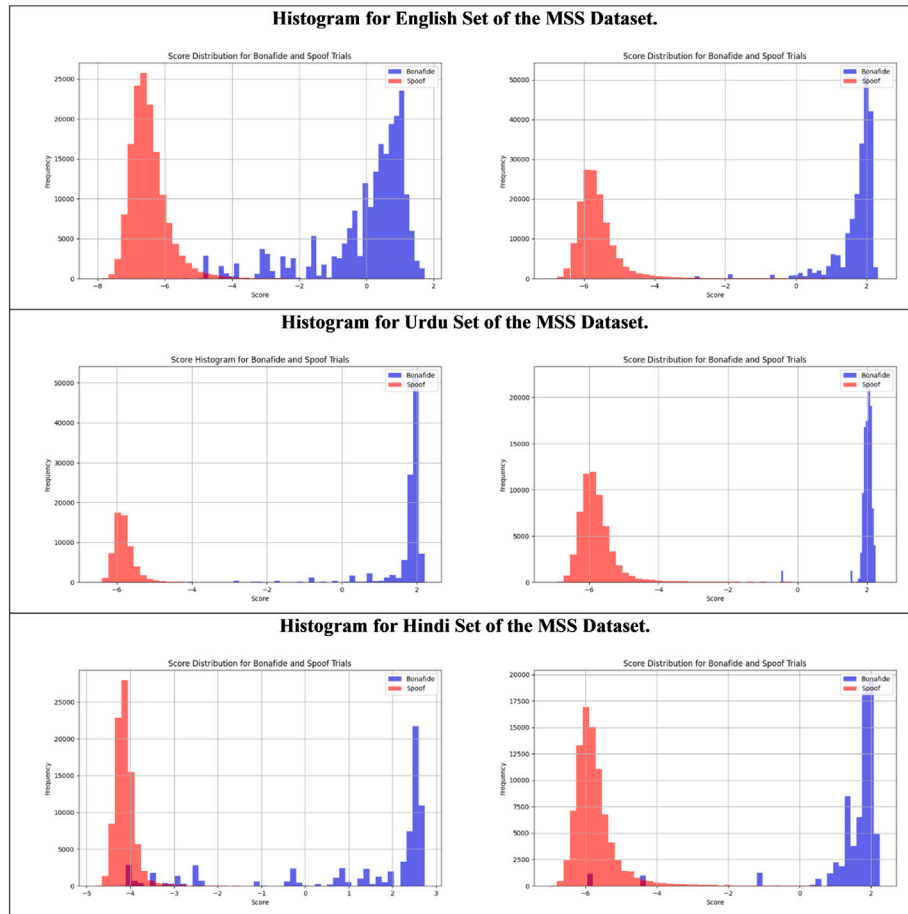


Fig. 4. Histograms of the scores for the bona fide and spoofed audio samples. Histograms on the left side are obtained when the proposed framework is trained and tested on the individual MSS language subsets. Histograms on the right side are obtained when the framework is trained on the combined set (Urdu, Hindi, and English) and tested on the individual language subsets of the MSS dataset.

To further validate the impact of multilingual training, we plotted histograms of score distributions for bona fide and spoofed samples under two scenarios. The first scenario involves training and testing the proposed model on individual English, Urdu, and Hindi subsets of the MSS dataset. The second scenario requires training on the combined Urdu–Hindi–English set and testing on the individual subsets. The resulting histograms, presented in Fig. 4, illustrate

the effect of training data diversity on the Wav2Vec2+UniSpeech-SAT framework's ability to discriminate between bona fide and spoofed utterances. As shown in Fig. 4, the bona fide samples exhibit score concentrations toward the positive range, whereas the spoofed samples cluster toward the negative range. When trained on individual language subsets, the histograms reveal noticeable overlap between the bona fide and spoofed score distributions, suggesting that the

Table 12
Comparison with existing audio spoofing detection methods.

Methods	ASV-19 LA		ASV-21 LA		ASV-21 DF
	min-tDCF	EER (%)	min-tDCF	EER (%)	EER (%)
WavLM – MLP Classifier (Tran et al., 2024)	–	0.23	–	3.31	4.47
Wav2Vec2 – RawNet2 (Kang et al., 2024)	0.0032	0.12	–	–	–
WavLM – Ecapa (Kulkarni et al., 2024)	0.022	0.80	0.3720	6.68	15.92
HuBERT – Ecapa (Kulkarni et al., 2024)	0.031	1.05	0.5397	12.55	13.79
Wav2Vec2 – LLGF (Wang and Yamagishi, 2021)	–	0.86	–	13.19	7.44
HuBERT – LLGF (Wang and Yamagishi, 2021)	–	3.55	–	9.55	13.07
Wav2Vec2 – Classifier (Martin-Doñas and Álvarez, 2022)	–	–	–	3.54	4.98
Wav2Vec2 – AASIST (Tak et al., 2022)	–	–	0.2099	0.82	2.85
WavLM – LSTM (Pan et al., 2024)	–	0.54	–	4.52	4.37
Wav2Vec2 – AASIST2 (Zhang et al., 2024)	–	0.15	–	1.61	2.77
Wav2Vec2 – Classifier (Li and Zhang, 2024)	0.0165	0.51	–	–	–
Wav2Vec2+UniSpeech-SAT	0.0094	0.38	0.2124	0.99	2.46
– AASIST (Proposed)					

model assigns similar scores to both classes, leading to a higher misclassification rate. In contrast, when the framework is trained on the combined multilingual dataset, the overlap between the two distributions is substantially reduced, indicating a lower misclassification rate and greater discrimination. Overall, this improvement demonstrates that the Wav2Vec2+UniSpeech-SAT model benefits from exposure to diverse linguistic characteristics, including acoustic and phonetic variations, during multilingual training. This diversity enables the model to better generalize across code-switched utterances from multiple languages, ultimately improving its performance in audio spoofing detection.

5.5. Comparison with existing methods

A comparative study was conducted to analyze the performance improvement achieved by the proposed framework. We specifically present a comparative analysis of the proposed Wav2Vec2+UniSpeech-SAT-based AASIST framework against the existing SSL-based SOTA audio spoofing methods on the ASV-19 LA, ASV-21 LA, and ASV-21 DF datasets. The comparison results are presented in Table 12. For the ASV-19 LA dataset, the Wav2Vec2 SSL model combined with RawNet2 achieved the best performance, with the lowest min-tDCF and EER, at 0.0032 and 0.12, respectively. However, our proposed framework achieves the second-lowest min-tDCF of 0.0094 on the ASV-19 LA dataset. Likewise, on the ASV-21 LA dataset, the proposed Wav2Vec2+UniSpeech-SAT framework achieved the second-best min-tDCF of 0.2124 and EER of 0.99. In addition, the proposed Wav2Vec2+UniSpeech-SAT framework outperforms all comparative methods on the ASV-21 DF dataset, achieving the lowest EER of 2.46, thus demonstrating its competitive effectiveness for detecting spoofed speech under the ASV-21 DF evaluation conditions.

We also compared the performance of the proposed framework with the baseline methods, namely RawNet2, AASIST, and Wav2Vec2 – AASIST, utilizing the proposed MSS dataset. Each model was trained with the same set of experimental protocols. The EERs are reported in Table 13. It can be observed that the proposed model outperforms the baseline methods on the MSS – English subset; its performance on MSS – Urdu is comparable to Wav2Vec2 – AASIST, whereas it is the second-best performer on MSS – Hindi. Moreover, the proposed model demonstrates good generalizability on the ASV-21 DF dataset (Table 12) compared to Wav2Vec2 – AASIST. Overall, the comparative results indicate that incorporating SSL representations into the AASIST framework can provide useful complementary information for code-switched speech spoof detection. However, the gains are not consistent across all language subsets, suggesting that the effectiveness of SSL-based frontends depends on the characteristics of the language and evaluation condition. Furthermore, the results demonstrate that the individual RawNet2 and AASIST with the Sinc layer as a frontend do not achieve better performance on the MSS dataset. These results suggest

that, under the evaluated conditions, the Sinc-based frontends are less effective than SSL-based representations for detecting code-switched spoofed audio, potentially due to the richer speech representations learned by the SSL models.

6. Conclusion

This paper presents the MSS dataset, comprising code-switched audio samples from Urdu and Hindi speakers across three languages: Urdu, Hindi, and English. In addition, a fused SSL-based AASIST framework is proposed, integrating Wav2Vec2 and UniSpeech-SAT frontends for robust audio spoofing detection. Extensive experiments on the proposed MSS and benchmark ASV-19 LA, ASV-21 LA, and ASV-21 DF datasets demonstrate the framework's effectiveness in detecting spoofed speech in both monolingual and code-switched audio samples. Attack-wise evaluation highlights the framework's strong performance against the challenging A17 attack in ASV-19 LA and identifies SeedVC as the most difficult spoofing attack in the MSS dataset. Comparative results further demonstrate the strong cross-corpus performance of the proposed framework on the ASV-21 DF dataset compared with SOTA methods. Cross-lingual and cross-corpora evaluations show that the model trained on the MSS dataset generalizes effectively across languages and corpora. In contrast, models trained on monolingual datasets (e.g., ASV-19 LA) fail to adapt to code-switched utterances. These findings emphasize the importance of multilingual, code-switched datasets in advancing robust spoofing detection. In future work, we aim to assess the robustness of SSL-based CMs against adversarial attacks and to develop explainable spoofing detection models to enhance transparency and trust in ASV systems.

CRedit authorship contribution statement

Hafsa Ilyas: Writing – original draft, Investigation, Formal analysis, Data curation. **Muhammad Hamza:** Methodology, Investigation, Formal analysis, Data curation. **Junaid Mir:** Writing – original draft, Methodology, Investigation, Conceptualization. **Ali Javed:** Supervision, Methodology, Investigation. **Muhammad Haroon Yousaf:** Validation, Resources, Project administration, Conceptualization. **Ahmed Zoha:** Resources, Project administration, Funding acquisition. **Shah Nawaz:** Writing – review & editing, Validation. **Habib Ullah Manzoor:** Writing – review & editing, Validation.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with language editing, proofreading, and improving the readability of the manuscript. After using this tool, the authors reviewed and edited the output as needed and take full responsibility for the content of the published article.

Table 13

Comparison with other baseline methods utilizing the MSS dataset. Results are reported in terms of EER (%).

Methods	MSS – Urdu	MSS – Hindi	MSS – English
RawNet2	5.47	12.28	6.69
AASIST	3.44	6.89	3.55
Wav2Vec2 – AASIST (Tak et al., 2022)	0.40	4.18	1.70
Wav2Vec2 + UniSpeech-SAT – AASIST (Proposed)	0.43	5.55	1.39

Funding

This work was supported by the Higher Education Commission, Pakistan [Grant No. 87309] and the Engineering and Physical Sciences Research Council (EPSRC), UK [Grant No. EP/Y002288/1].

Declaration of competing interest

- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article submitted to Engineering Applications of Artificial Intelligence.
- The work described has not been published previously.
- The article is not under consideration for publication elsewhere.
- The article's publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out.
- If accepted, the article will not be published elsewhere in the same form, in English or in any other language, including electronically, without the written consent of the copyright-holder.

Data availability

Data will be made available on request.

References

- Al-Tairi, H., et al., 2025. Deepplad countermeasure for logical access audio spoofing. *Sci. Rep.* 15 (1), 20839.
- Ali, H., et al., 2025. Multilingual dataset integration strategies for robust audio deepfake detection: A SAFE challenge system. In: *IEEE Automatic Speech Recognition and Understanding Workshop. ASRU*.
- Arif, T., et al., 2021. Voice spoofing countermeasure for logical access attacks detection. *IEEE Access* 9, 162857–162868.
- ASVspoof Challenge, (2025). <https://www.asvspoof.org/>. (Accessed on 25 February 2025).
- Baas, M., van Niekerk, B., Kamper, H., 2023. Voice conversion with just nearest neighbors. *arXiv preprint arXiv:2305.18975*.
- Baevski, A., et al., 2020a. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460.
- Baevski, A., et al., 2020b. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *NeurIPS 2020*.
- Chen, S., et al., 2022a. UniSpeech-SAT: Universal speech representation learning with speaker aware pre-training. In: *ICASSP 2022. IEEE*.
- Chen, S., et al., 2022b. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* 16 (6), 1505–1518.
- Choi, H.-Y., Lee, S.-H., Lee, S.-W., 2023. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. *arXiv preprint arXiv:2311.04693*.
- Ciobanu, I.-P., et al., 2026. XMAD-bench: Cross-domain multilingual audio deepfake benchmark. In: *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3109–3120.
- Das, A., et al., 2025. Generalizable audio spoofing detection using non-semantic representations. *arXiv preprint arXiv:2509.00186*.
- Dawood, H., et al., 2022. A robust voice spoofing detection system using novel CLS-lbp features and LSTM. *J. King Saud Univ.-Comput. Inf. Sci.* 34 (9), 7300–7312.
- Di Pierno, A., et al., 2025. End-to-end audio deepfake detection from raw waveforms: A RawNet-based approach with cross-dataset evaluation. In: *International Joint Conference on Neural Networks. IJCNN, IEEE*.
- Farooq, M.U., et al., 2026. TRACE: Training-free partial audio deepfake detection via embedding trajectory analysis of speech foundation models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR*.
- Fathan, A., Alam, J., Kang, W.H., 2022. Mel-spectrogram image-based end-to-end audio deepfake detection under channel-mismatched conditions. In: *ICME 2022. IEEE*.
- Febrinanto, F.G., et al., 2025. Vision graph non-contrastive learning for audio deepfake detection with limited labels. *ArXiv E-Prints*.
- Flórez, P.T., Manrique, R., Nunes, B.P., 2023. HABLE: A dataset of latin American spanish accents for voice anti-spoofing. In: *INTERSPEECH 2023*.
- Frank, J., Schönherr, L., 2021. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*.
- Goel, C., et al., 2024. Towards attention-based contrastive learning for audio spoof detection. *arXiv preprint arXiv:2407.03514*.
- Hamza, A., et al., 2022. Deepfake audio detection via MFCC features using machine learning. *IEEE Access* 10, 134018–134028.
- Hsu, W.-N., et al., 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 3451–3460.
- Ilyas, H., Javed, A., Malik, K.M., 2023. AvFakeNet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Appl. Soft Comput.* 136, 110124.
- Javed, A., et al., 2021. Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. *Appl. Acoust.* 183, 108283.
- Javed, A., et al., 2022. Voice spoofing detector: A unified anti-spoofing framework. *Expert Syst. Appl.* 198, 116770.
- weon Jung, J., et al., 2019. RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. *arXiv preprint arXiv:1904.08104*.
- weon Jung, J., et al., 2022. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In: *ICASSP 2022. IEEE*.
- Kang, T., et al., 2024. Experimental study: Enhancing voice spoofing detection models with wav2vec 2.0. *arXiv preprint arXiv:2402.17127*.
- Kheir, Y.E., et al., 2025. Comprehensive layer-wise analysis of SSL models for audio deepfake detection. *arXiv preprint arXiv:2502.03559*.
- Kinnunen, T., et al., 2017. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.
- Kulkarni, A., et al., 2024. Exploring generalization to unseen audio data for spoofing: Insights from SSL models. In: *ASVspoof Workshop*.
- Le, T.D.N., Teh, K.K., Tran, H.D., 2024. Continuous learning of transformer-based audio deepfake detection. *ArXiv E-Prints*.
- Lee, S.-H., et al., 2023. HierSpeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv preprint arXiv:2311.12454*.
- Li, J., Tu, W., Xiao, L., 2023. Freevc: Towards high-quality text-free one-shot voice conversion. In: *ICASSP 2023. IEEE*.
- Li, M., Zhang, X.-P., 2024. Interpretable temporal class activation representation for audio spoofing detection. *arXiv preprint arXiv:2406.08825*.
- Liu, S., 2024. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*.
- Liu, T., et al., 2025. Nes2Net: A lightweight nested architecture for foundation model driven speech anti-spoofing. *IEEE Trans. Inf. Forensics Secur.* 20, 12005–12018.
- Ma, H., et al., 2024. CFAD: A Chinese dataset for fake audio detection. *Speech Commun.* 164, 103122.
- Martin-Doñas, J.M., Álvarez, A., 2022. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 ADD challenge. In: *ICASSP 2022. IEEE*.
- Mawalim, C.O., et al., 2026. Multilingual deepfake speech dataset for robust and generalizable detection. *IEEE Access* 14, 57144–57161.
- Müller, N.M., et al., 2022. Does audio deepfake detection generalize?. *arXiv preprint arXiv:2203.16263*.
- Müller, N.M., et al., 2024. MLAAD: The multi-language audio anti-spoofing dataset. In: *IJCNN 2024. IEEE*.
- Nawaz, S., Saeed, M.S., Morerio, P., Mahmood, A., Gallo, I., Yousaf, M.H., Del Bue, A., 2021. Cross-modal speaker verification and recognition: A multilingual perspective. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, IEEE*, pp. 1682–1691.
- Pan, Z., et al., 2024. Attentive merging of hidden embeddings from pre-trained speech model for anti-spoofing detection. *arXiv preprint arXiv:2406.10283*.
- Rahmeni, R., Aicha, A.B., Ayed, Y.B., 2022. Voice spoofing detection based on acoustic and glottal flow features using conventional machine learning techniques. *Multimedia Tools Appl.* 81 (22), 31443–31467.
- Ranjan, R., Vatsa, M., Singh, R., 2022. Statnet: Spectral and temporal features based multi-task network for audio spoofing detection. In: *IEEE International Joint Conference on Biometrics. IJCB, IEEE*.

- Reimao, R., Tzerpos, V., 2019. For: A dataset for synthetic speech detection. In: ICST 2019. IEEE.
- Salvi, D., et al., 2023. TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection. *IEEE Access* 11, 50851–50866.
- Sinha, S., Dey, S., Saha, G., 2024. Improving self-supervised learning model for audio spoofing detection with layer-conditioned embedding fusion. *Comput. Speech Lang.* 86, 101599.
- Sun, C., et al., 2023. AI-synthesized voice detection using neural vocoder artifacts. In: CVPR 2023.
- Sun, J., et al., 2025. Contrastive learning-based speech spoofing detection for multimedia security in edge intelligence. *ACM Trans. Multimed. Comput. Commun. Appl.* 21 (8), 1–21.
- Tak, H., et al., 2021a. End-to-end anti-spoofing with RawNet2. In: ICASSP 2021. IEEE.
- Tak, H., et al., 2021b. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710*.
- Tak, H., et al., 2021c. Graph attention networks for anti-spoofing. *arXiv preprint arXiv:2104.03654*.
- Tak, H., et al., 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*.
- Tran, H.M., et al., 2024. Spoofed speech detection with a focus on speaker embedding. In: INTERSPEECH.
- Viakhirev, I., et al., 2025. Towards scalable AASIST: refining graph attention for speech deepfake detection. *arXiv preprint arXiv:2507.11777*.
- Wang, Z., Hansen, J.H., 2024. Towards improving synthetic audio spoofing detection robustness via meta-learning and disentangled training with adversarial examples. *IEEE Access*.
- Wang, X., Yamagishi, J., 2021. Investigating self-supervised front ends for speech spoofing countermeasures. *arXiv preprint arXiv:2111.07725*.
- Wang, X., et al., 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* 64, 101114.
- Wang, X., et al., 2023. Robust audio anti-spoofing countermeasure with joint training of front-end and back-end models. In: Interspeech 2023.
- Wang, X., et al., 2024. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739*.
- Wu, Z., et al., 2015. Asvspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In: INTERSPEECH 2015. ISCA.
- Yamagishi, J., et al., 2021. Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.
- Yi, J., et al., 2022. ADD 2022: The first audio deep synthesis detection challenge. In: ICASSP 2022. IEEE.
- Zaman, K., et al., 2026. Deepfake audio detection using self-supervised fusion representations. *arXiv preprint arXiv:2605.03420*.
- Zhang, Z., et al., 2021. FMFCC-a: A challenging mandarin dataset for synthetic speech detection. In: International Workshop on Digital Watermarking. Springer.
- Zhang, Y., et al., 2024. Improving short utterance anti-spoofing with AASIST2. In: ICASSP 2024. IEEE.