Int. J. Anal. Appl. (2026), 24:43

International Journal of Analysis and Applications

Cross-Lingual Speech Emotion Recognition with Attention-Driven Bi-LSTM: Advancing Kashmiri and Multilingual Adaptation

GH Mohmad Dar[1], Radhakrishnan Delhibabu[2,*]

[1]*Department of Mathematics, School of Advanced Sciences,Vellore Institute of Technology, Vellore, Tamil Nadu, India*

[2]*Vellore Institute of Technology, Vellore, Tamil Nadu, India*

[*]*Corresponding author:* rdelhibabu@vit.ac.in

Abstract. Speech Emotion Recognition (SER) has achieved notable success in high-resource languages, yet remains underexplored for Kashmiri, a low-resource Dardic language characterized by tonal and prosodic complexity. This study introduces the first systematic framework for Kashmiri SER and examines its cross-lingual adaptability using Urdu, Persian, and English language datasets. A Bidirectional Long Short-Term Memory (Bi-LSTM) network with attention mechanism was employed to capture bidirectional temporal dependencies while emphasizing emotionally salient segments, with Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram features as inputs. Three experiments were conducted: within-language evaluation yielded high accuracies (93.2% for Kashmiri, 97% for Urdu, 85% for Persian, and 80.05% for English); cross-lingual transfer revealed substantial performance decline (25–34%), highlighting phonetic and prosodic mismatches; and progressive domain adaptation improved results up to 89%, 81%, and 83% for Urdu, Persian, and English, respectively. These findings demonstrate the challenges of direct transfer and the promise of adaptation, offering a pathway toward resource-efficient, multilingual SER for underrepresented languages.

1. Introduction

Speech is one of the most intuitive and powerful modes of human communication, and emotions shape both the production and perception of speech[1]. Understanding the emotional state of a speaker is valuable for a variety of real-world applications, including mental health monitoring, driver assistance systems, call center analytics, education technology, and entertainment systems [2]. Speech Emotion Recognition (SER) aims to automatically classify emotions such as anger, sadness, happiness, and neutrality from acoustic signals, independent of linguistic content [3]. With the increasing prevalence of voice-driven interfaces, robust SER systems capable of

Received: Oct. 18, 2025.

2020 *Mathematics Subject Classification.* 68T10.

Key words and phrases. speech emotion recognition (SER); Mel-frequency cepstral coefficients (MFCC); Kashmiri language; long short-term memory (LSTM); cross lingual adaptation; attention mechanism.

https://doi.org/10.28924/2291-8639-24-2026-43
ISSN: 2291-8639

© 2026 the author(s).

functioning across languages and speaker populations are becoming an essential component of human–computer interaction systems[4][5].

Significant progress in SER has been made in the past decade, driven by the availability of public datasets and advances in machine learning architectures such as convolutional neural networks (CNNs)[6], recurrent neural networks (RNNs), and transformers[7]. These models typically rely on well-known acoustic features such as Mel-frequency cepstral coefficients (MFCCs), prosodic features, and spectrogram representations [8][9], which are well-suited for capturing the spectral and temporal dynamics of speech. High-resource languages such as English, German, and Mandarin benefit from large annotated emotion corpora like IEMOCAP, RAVDESS, and Emo-DB, enabling data-intensive deep models to achieve state-of-the-art performance [10].

However, the situation is very different for low-resource languages. Kashmiri a Dardic language spoken by over seven million people, mainly in Jammu and Kashmir (India) and parts of Pakistan is severely underrepresented in computational speech research [11]. To date, there is no large-scale publicly available emotion-labeled corpus for Kashmiri, and there are very few studies investigating SER for this language [12]. This lack of data leads to poor generalization when training models solely on Kashmiri, as deep learning models require substantial labeled samples to learn robust patterns. Moreover, Kashmiri exhibits unique phonetic and prosodic patterns (such as its vowel inventory and tonal variations) that are not adequately captured by models trained only on unrelated languages, thus limiting zero-shot transferability[13].

A promising solution is cross-lingual learning, which transfers knowledge from high-resource languages to low-resource target languages. Emotions often share universal acoustic correlates for example, anger is associated with higher energy and faster speech, whereas sadness is marked by lower intensity and slower tempo [14]. Exploiting these shared cues allows models trained on multiple source languages to improve recognition performance in low-resource languages. Recent works have explored approaches such as domain adaptation, adversarial learning, and transfer learning from multilingual speech representations [15][16], demonstrating that combining data from multiple languages leads to better cross-lingual generalization. Nevertheless, two critical gaps remain:

- Most cross-lingual SER studies adopt a simple source target paradigm, training on a single language and testing on another, without investigating how systematically increasing the proportion of auxiliary language data influences target-language performance.
- There is virtually no quantitative analysis for Kashmiri, meaning that researchers lack empirical evidence on how much multilingual data is required to achieve acceptable performance on this low-resource language.

This study addresses these gaps by conducting a comprehensive quantitative analysis of cross-lingual training for Kashmiri SER. We extract acoustic features using MFCCs and spectrograms to represent complementary spectral–temporal information, and we employ a bidirectional long short-term memory (Bi-LSTM) network to model temporal dependencies across

speech frames. Our experimental design systematically increases the proportion of auxiliary data from Urdu, Persian, and English during training and evaluates the resulting performance on Kashmiri. The results reveal a clear trend: gradually incorporating more data from related and high-resource languages significantly improves recognition accuracy, highlighting the importance of multilingual data augmentation strategies for low-resource SER.

By presenting one of the first systematic studies on Kashmiri speech emotion recognition and offering a quantitative analysis of cross-lingual data effects, this work provides critical insights for building practical, data-efficient, and language-inclusive SER systems. The findings can guide researchers and practitioners in curating training corpora for other underrepresented languages facing similar challenges.

1.1. **Aims and Contributions.** The overarching aim of this research is to advance speech emotion recognition (SER) in low-resource languages, with a particular focus on Kashmiri, by leveraging cross-lingual transfer learning strategies. The study first establishes a strong baseline for Kashmiri SER using complementary acoustic features, namely Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, modeled through a Bidirectional Long Short-Term Memory (Bi-LSTM) network with attention. Building upon this foundation, the work investigates the generalization capability of Kashmiri-trained models to other languages (Urdu, Persian, and English) and proposes a progressive domain adaptation strategy where incremental proportions of target language data are integrated into training. This design enables a systematic evaluation of how multilingual knowledge transfer can mitigate data scarcity while preserving model efficiency.

The major contributions of this work are outlined next.

- We present the first empirical baseline for cross lingual Speech Emotion Recognition (SER) in Kashmiri, a critically underrepresented Dardic language, by developing and evaluating a curated emotional speech corpus. This establishes a foundation for future research in low-resource languages.
- We propose a structured cross-lingual transfer and adaptation framework, including zero-shot transfer (Kashmiri $\rightarrow$ Urdu, Persian, English) and a progressive domain adaptation strategy, to quantify how incremental inclusion of target data improves recognition performance.
- We integrate Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram features within a Bi-LSTM attention framework, enabling the joint modeling of spectral and temporal dynamics and yielding robust, discriminative emotion representations.
- We demonstrate data-efficient knowledge transfer by identifying the minimal proportion of auxiliary data required to substantially enhance cross-lingual SER, offering practical guidelines for resource-constrained multilingual applications.
- We establish a generalizable experimental paradigm that is replicable across other low-resource and typologically diverse languages, advancing the development of inclusive and multilingual SER systems.

2. Realted study:

Speech Emotion Recognition (SER) plays a vital role in Human-Computer Interaction by enabling systems to detect and interpret human emotions from speech. While significant progress has been made for high-resource languages using deep learning and well-annotated corpora, low-resource languages remain largely unexplored due to limited data and unique phonetic and prosodic characteristics. Cross-lingual SER, which leverages annotated data from high-resource languages, provides a promising solution by exploiting universal acoustic patterns, such as higher pitch for anger and slower, low-energy speech for sadness.

In this section, we review prior work relevant to SER in low-resource languages focusing on dataset development, preprocessing , feature extraction techniques and classification approaches. We then discuss cross-lingual SER strategies, highlighting how knowledge transfer from multiple languages can improve emotion recognition in underrepresented languages such as Kashmiri.

2.1. **Speech Emotion Recognition in Underexplored Languages.**

In recent years, several emotional speech databases have been created to support research in Speech Emotion Recognition (SER) for underrepresented Indo-Aryan and Dravidian languages. One of the major resources is the Hindi emotional speech corpus (IITKGP-SEHSC)[17], which includes 12,000 speech samples covering eight emotions such as happiness, sadness, anger, sarcasm, and fear. For Urdu, publicly available datasets contain 400 utterances from 38 speakers, as well as a larger dataset developed by Asghar et al.[18] with 2,500 utterances across five emotions. In Bangla, the SUBESCO dataset is the largest to date, offering more than 7,000 utterances and covering emotions like anger, sadness, fear, and surprise. Other efforts include emotional datasets for Oriya (developed by Mohanty and Swain) [19], Assamese (which includes five native dialects like Bodo and Mishing), and Punjabi (with 900 utterances collected by Kaur and Singh). In the Dravidian language group, Kannada and Malayalam emotional speech corpora have also been developed, with recordings of emotions such as happiness, anger, sadness, fear, and neutral states[20, 21]. These datasets show increasing interest in extending SER research to India's regional languages.

Along with data collection, research methods in SER for these languages have evolved. Earlier approaches mostly used traditional machine learning models with features like LPCC, pitch, and energy. More recent studies use advanced deep learning techniques, which have shown better accuracy. For example, a study by Agarwal and Om [22] achieved 93.75% accuracy on Hindi data using a deep neural network optimized with a metaheuristic algorithm. Another study by Sultana et al.[23] combined convolutional and recurrent neural networks (DCNN and BLSTM) to achieve over 80% accuracy on Bangla and English datasets. Swain et al.[19] developed a deep ensemble model for the Oriya dataset, which performed better than previous models. These results suggest that deep learning methods, combined with good quality data, can significantly improve SER performance in low-resource languages.Table 1 summarizes some works on speech emotion recognition for Indo-Aryan and Dravidian languages.

TABLE 1. Summary of SER Studies in Indo-Aryan and Dravidian Languages

S.No	Reference	Dataset	Language(s)	Type	Emotions Covered
1	Syed et al., 2020 [24]	Urdu-Sindhi speech emotion corpus	Urdu, Sindhi	Audio	Happiness, anger, sadness, disgust, surprise, sarcasm, neutral
2	Samantaray et al., 2015 [25]	MESDNEI	Assamese	Audio	Happy, anger, fear, disgust, surprise, sad, neutral
3	Ali et al., 2015 [26]	Speech emotional corpus	Urdu, Sindhi, Pashto, Punjabi, Balochi	Audio	Happiness, sad, anger, neutral
4	Darekar et al., 2018 [27]	Marathi database	Marathi	Audio	Happy, sad, angry, fear, neutral, surprised
5	Basu et al., 2020 [28]	Santali Speech Data	Santali	Audio	Anger, fear, happy, sad, surprise, neutral
6	Dhar et al., 2021 [29]	Bangla emotional speech dataset	Bangla	Audio	Happy, angry, neutral
7	Fernandes et al., 2021 [30]	Hindi emotional speech database	Hindi	Audio	Happy, sad, anger, fear, surprise, disgust, neutral
8	Tank et al., 2020 [31]	Gujarati speech database	Gujarati	Audio	Sadness, surprise, anger, disgust, fear, happiness
9	Aziz et al., 2023 [32]	BanglaSER	Bengali	Audio	Anger, Disgust, Happiness, Sadness, Surprise, Fear, Neutral

2.2. **Preprocessing Techniques and Feature Selection in SER.** Speech Emotion Recognition (SER) aims to analyze vocal expressions to determine a speaker's emotional state. Two essential components of any SER pipeline are *preprocessing* and *feature selection*, both of which significantly influence the system's accuracy and robustness. However, the design and implementation of these components are not standardized across studies, as they often vary with language characteristics, dataset types, and modeling strategies.

Preprocessing plays a vital role in improving the quality of the speech signal before feature extraction. Common methods include silence removal, noise reduction, and signal normalization [33–35]. These techniques help reduce irrelevant information and variability in the signal, ensuring that only emotion-relevant acoustic features are retained. This becomes especially crucial when dealing with naturalistic or in-the-wild recordings where background noise and inconsistent recording setups can affect signal quality.

Interestingly, some recent works skip extensive preprocessing, especially when employing spectrogram or mel-spectrogram images as input [36]. The rationale is that preprocessing steps like filtering or silence removal may eliminate subtle acoustic cues essential for accurate emotion recognition. In such cases, minimal preprocessing such as trimming silent portions at the start and end of an utterance or applying light denoising may be preferred to preserve emotional nuances [35].

For systems relying on traditional feature extraction, more elaborate preprocessing steps are adopted. These include techniques such as framing, windowing, pre-emphasis, and endpoint detection [37, 38]. These methods allow the speech signal to be segmented into small, analyzable time frames, often enhancing specific frequency bands to better capture emotional content. This is particularly useful when deriving handcrafted features like MFCCs, LPCs, pitch, energy, or formant frequencies.

In terms of feature selection, SER research commonly categorizes features into three types: prosodic (e.g., pitch, energy), spectral (e.g., MFCCs, spectral flux), and voice quality features (e.g., jitter, shimmer). Some works also employ higher-order statistical descriptors or hyper-prosodic features to summarize temporal dynamics [39]. These descriptors often provide valuable information about patterns across longer time spans and improve model performance when used alongside traditional features.

Among these, Mel-Frequency Cepstral Coefficients (MFCCs) have remained a dominant feature set due to their efficiency in capturing the spectral properties of speech signals. First introduced by Davis and Mermelstein [40], MFCCs have been extensively studied and optimized for robustness in noisy environments [41]. Research by Oflazoglu et al. [42] demonstrated that MFCCs are sensitive to changes in emotional intensity, thereby making them suitable for fine-grained emotion analysis. Moreover, Sarma et al. [43] successfully used MFCCs for emotion classification across multiple languages, showing their adaptability even in resource-scarce linguistic contexts.

Given that our study explores an underrepresented language like Kashmiri, the choice of preprocessing and features becomes even more critical. Languages differ in their phonetic and prosodic structures, which directly influences how emotion is expressed and perceived. Therefore, the preprocessing pipeline must be carefully adapted to preserve emotionally salient patterns unique to Kashmiri speech. Additionally, robust feature sets such as MFCCs offer a promising direction for building effective emotion recognition systems in such linguistically unexplored settings.Table 2 illustrates Preprocessing Techniques and Features Used in Recent SER Studies.

2.3. **Deep Learning Classifiers in SER.** Following the stages of preprocessing and acoustic feature extraction, the classification component of a Speech Emotion Recognition (SER) system serves as the core mechanism for identifying emotional states from speech patterns. In recent years, deep learning architectures have demonstrated substantial performance gains over traditional machine learning techniques due to their capacity to model complex and non-linear feature relationships, particularly in sequential and high-dimensional data.

Table 2. Preprocessing Techniques and Features Used in Recent SER Studies

S.No	Reference	Preprocessing Techniques	Features Used
1	Mohanty et al., 2024 [44]	Z-normalization, Silence removal	MFCC, Spectral Entropy, Chroma Features
2	Chowdhury et al., 2025 [45]	Pre-emphasis, Noise reduction	MFCC, ZCR, Chroma, STFT
3	Leem et al., 2023 [46]	Min-Max Normalization, Segmentation	Acoustic Features
4	Yuanchao et al., 2023 [47]	Data augmentation (pitch shifting), Normalization	Mel-spectrogram, Pitch, Intensity
5	Kaur et al., 2022 [48]	Noise addition, Time warping	MFCC, Spectral Contrast, Tonnetz
6	Sajjad et al., 2020 [35]	K-means clustering, STFT	Spectrogram-based CNN features

Among these architectures, Recurrent Neural Networks (RNNs) and their variants, especially Long Short-Term Memory (LSTM) networks, have become the standard for temporal modeling in SER tasks. LSTMs were introduced by Hochreiter and Schmidhuber [49] to address the vanishing gradient problem inherent in standard RNNs, enabling the learning of long-range dependencies within time-series data. Their internal memory and gating mechanisms allow for selective retention and forgetting of temporal information, a critical property for modeling the dynamic progression of affective cues in speech.

Graves [50] empirically validated the superiority of LSTMs for frame-wise sequence modeling, demonstrating their efficacy in contexts where the emotional content unfolds gradually over time. Additionally, modifications by Gers et al. [51] enhanced the memory management mechanisms of LSTMs, thereby increasing their sensitivity to salient prosodic variations.

Recent advances have led to hybrid models that combine Convolutional Neural Networks (CNNs) with LSTM units. CNNs are proficient in capturing local spectral patterns and spatial hierarchies from 2D inputs such as spectrograms, while LSTMs process these features temporally to capture the evolution of emotion across the utterance. Such hybrid CNN-LSTM architectures have shown improved recognition accuracy across several SER benchmarks [52], especially in scenarios involving mel-spectrogram or log-filterbank representations.

These models are particularly advantageous when working with underrepresented and low-resource languages, such as Kashmiri. Due to the scarcity of annotated data and the variability in emotional expression across languages, models that can generalize across heterogeneous data distributions are essential. The ability of deep neural networks to extract invariant and transferable feature representations makes them well-suited for cross-linguistic emotion recognition tasks.

Furthermore, classification performance in SER is closely tied to the granularity of temporal segmentation, particularly the choice of frame size. Turgut et al. [53] demonstrated that optimal frame length significantly influences the sensitivity of the classifier to fine-grained emotional variations. As such, the interaction between frame resolution and classifier architecture must be carefully optimized to ensure reliable emotion recognition across diverse linguistic and acoustic conditions. Table 3 illustrates how the progress has been made in emotion detection, focusing on how different models and techniques can be used to improve emotion detection in different languages, such as Kashmiri.

2.4. **Cross-Lingual Speech Emotion Recognition.** Cross-lingual SER addresses the challenge of limited annotated data in low-resource languages by transferring knowledge from high-resource languages. Emotions often share universal acoustic cues, such as elevated pitch and faster articulation for anger or reduced energy and slower delivery for sadness. Recent studies have shown that cross-lingual representations can substantially improve performance by enabling models to generalize across typologically distinct languages. For example, Yang et al. [63] proposed WavLM-based domain emotion embeddings that achieved strong single- and cross-lingual recognition results, while Zaidi et al. [64] introduced a multimodal dual-attention transformer framework that effectively transfers emotional knowledge with minimal target language data. These findings confirm that transfer learning and multilingual modeling can mitigate data scarcity, improving robustness and adaptability of SER systems in underexplored linguistic contexts such as Kashmiri.

A key insight from prior studies is that while cross-lingual transfer is promising, its effectiveness often diminishes for typologically distant languages due to differences in phonetic and prosodic realizations of emotion. Most research to date has focused on high-resource languages such as German, English, and Mandarin, leaving low-resource languages like Kashmiri largely unexplored. This gap motivates our study, which systematically investigates cross-lingual adaptation for Kashmiri in relation to Urdu, Persian, and English, thereby extending SER research into an underrepresented language family and highlighting the broader potential of multilingual knowledge transfer.

3. Methodology

This study presents a systematic framework for developing and evaluating a cross-lingual Speech Emotion Recognition (SER) system, with Kashmiri as the primary focus. As one of the least explored Dardic languages, Kashmiri faces severe data scarcity, making it an ideal testbed for

Table 3. Summary of Related Studies in SER and Model Methodologies

S.No	Reference	Focus	Methodology	Key Findings
1	Ntalampiras et al., 2011 [54]	Impact of frame size on emotion detection models	Varying frame sizes	Demonstrated that optimal frame size selection is crucial for achieving high classification accuracy
2	Bai et al., 2018 [55]	Sequence modeling alternatives	Temporal Convolutional Networks (TCNs)	TCNs offer stable gradients and parallel processing, outperforming traditional RNNs in sequence modeling
3	Cho et al., 2014 [56]	Balancing complexity and performance in emotion recognition	Gated Recurrent Units (GRUs)	GRUs are computationally efficient and effective in modeling emotional dynamics in speech
4	Kim et al., 2017 [57]	Feature extraction and hierarchical representation	Deep Convolutional Neural Networks (DCNNs)	DCNNs efficiently capture both local and global acoustic patterns in audio signals
5	Shen et al., 2018 [58]	Addressing data scarcity in underrepresented languages	Transfer learning, hybrid models	Showed that hybrid deep learning models can adapt well to low-resource language scenarios
6	Su et al., 2020 [59]	Emotion detection from speech	Genetic Algorithm optimized GRU (GA-GRU)	Outperformed existing models in utterance-level emotion recognition across datasets
7	Manohar et al., 2022 [60]	Human emotion recognition	LSTM and CNN hybrid networks	Demonstrated enhanced performance by leveraging both spatial and temporal features in audio
8	Sultana et al., 2022 [61]	Bangla speech emotion recognition	Deep CNN and Bidirectional LSTM (BLSTM) networks	Successfully implemented cross-lingual SER using deep CNN and BLSTM models
9	Wang et al., 2023 [62]	Emotion detection in low-resource languages	Bi-LSTM, GRU, DCNN, TCN	Highlighted the role of model selection and frame size in emotion detection for resource-scarce languages

low-resource SER research. To address this, we curated a Kashmiri emotional speech dataset and integrated it with auxiliary resources from Urdu, Persian, and English. This multilingual setup enables us to investigate both within-language baselines and cross-lingual knowledge transfer. Figure 1 illustrates the overall methodological pipeline.

TABLE 4. Summary of Cross-Lingual SER Studies: Methods, Features, Databases, and Experimental Focus

S.No	Reference	Methods Employed	Feature Extraction	Databases (Languages)	Expts.
1	Yang et al. (2024)	WavLM embeddings + multi-task learning	WavLM, acoustic embeddings	Multiple (cross-lingual)	Single- and cross-lingual
2	Zaidi, Latif, Qadir (2023)	Multimodal dual-attention transformer (MDAT)	Spectral + multimodal features	Cross-lingual corpora	Cross-lingual

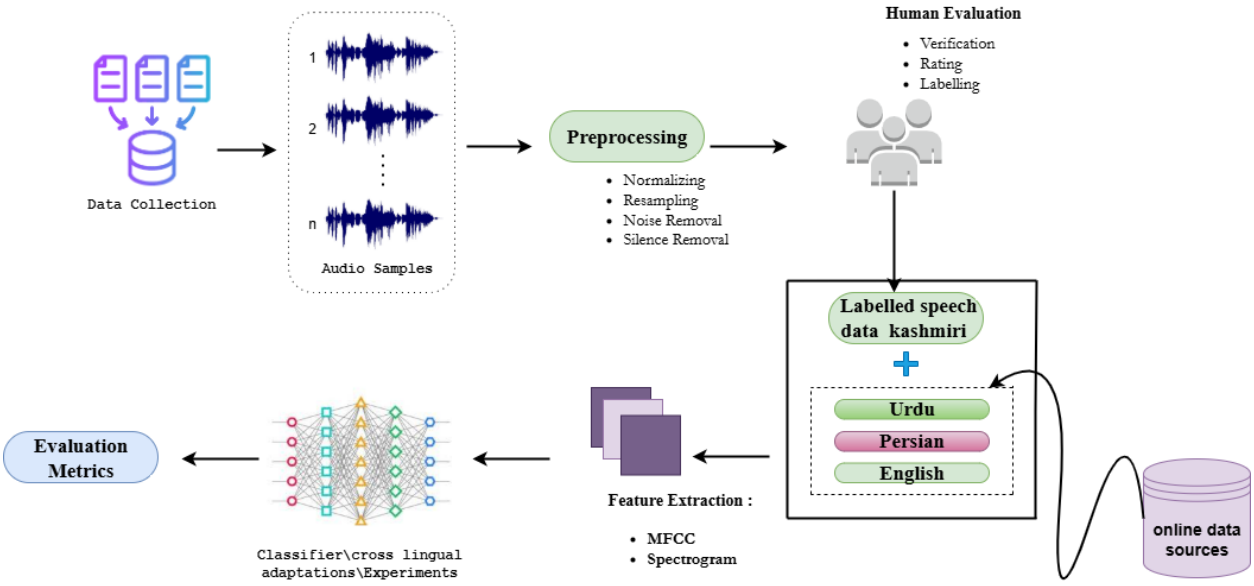


FIGURE 1. Overview of the proposed methodology for cross-lingual speech emotion recognition.

3.1. **Target Language Data: Kashmiri Speech Collection.** The absence of annotated emotional speech corpora for Kashmiri presents a key barrier to advancing SER in this language. To address this, we designed a rigorous data collection and annotation framework that emphasizes linguistic authenticity, demographic diversity, and high-quality labeling.

3.1.1. *Data Collection.* The corpus was created from 166 native Kashmiri speakers (88 male, 78 female) representing different age groups and regional dialects. Participants were drawn from northern, southern, and central regions of Kashmir to capture dialectal variation and ensure generalizability. The final dataset spans approximately 5 hours and 30 minutes of speech.

To maintain realism, 80% of the data consists of natural emotional speech, collected through semi-structured dialogues and storytelling tasks that elicited spontaneous emotions such as anger, happiness, sadness, and excitement. Controlled recordings (10%) were also included, where participants read phonetically balanced sentences designed by linguists and cultural experts. These acted samples ensured balance across emotion classes. The remaining 10% was sourced from ethically permissible public material, such as interviews, podcasts, and YouTube content, providing diversity in speaking styles, environments, and emotional intensities

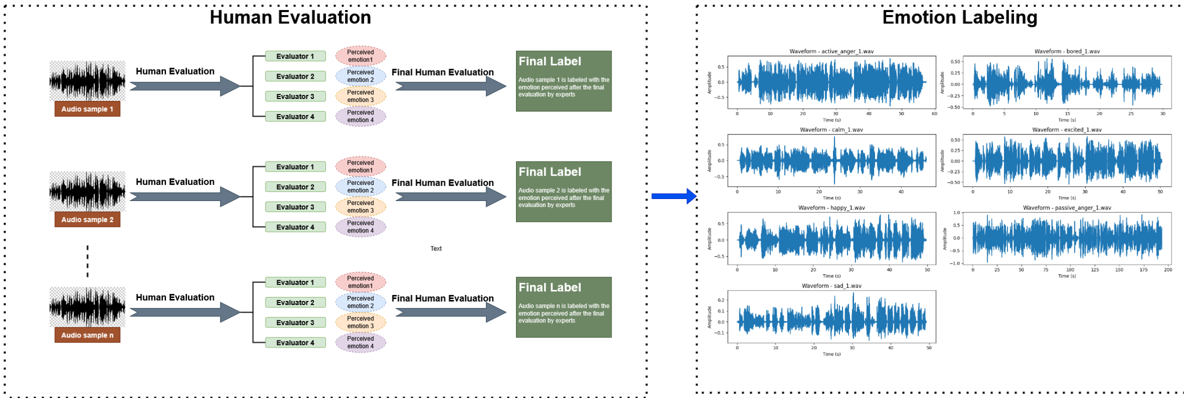
3.2. **Recording Setup and Quality Control.** All recordings were captured using studio-grade condenser microphones at 16 kHz, 16-bit resolution, and stored in mono-channel WAV format. Controlled sessions were held in acoustically treated environments, following a standardized protocol for microphone placement, noise control, and calibration. This ensured uniformity across sessions and minimized channel variability, producing clean, high-quality speech essential for robust SER.

3.3. **Raw Audio Preprocessing Pipeline.** The collected audio underwent a structured preprocessing pipeline to standardize quality. Noise reduction was applied using adaptive filtering and spectral subtraction. Peak normalization ensured consistent loudness, while silence trimming removed non-speech intervals. All recordings were saved as 16 kHz, 16-bit mono WAV files, ensuring compatibility and reproducibility for feature extraction tasks such as MFCCs and spectrograms. This pipeline was uniformly applied across Kashmiri and auxiliary corpora to maintain consistency for cross-lingual experiments.

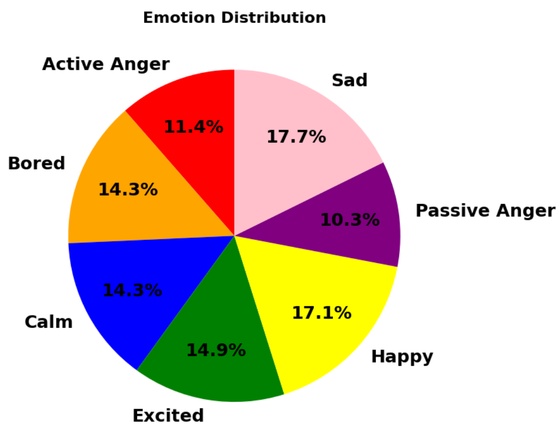
3.4. **Human Evaluation and Labeling.** Emotion annotation was carried out using a two-stage human evaluation process designed to minimize subjectivity. Audio samples were first annotated via Google Forms by four native Kashmiri raters with diverse socio-cultural and dialectal backgrounds. Raters were asked to: (i) identify the expressed emotion, (ii) justify their decision by pointing to prosodic cues (tone, pitch, intensity), and (iii) rate the intensity on a 1–5 scale.

Labels were assigned through majority voting where agreement was strong. In cases of disagreement particularly for overlapping categories such as happiness vs. excitement an adjudication panel of linguists and SER experts reviewed the cases and finalized the labels by consensus. This procedure ensured both perceptual validity and cultural appropriateness. Figure 2, summarize the annotation workflow, emotion distribution, and representative labeled samples.

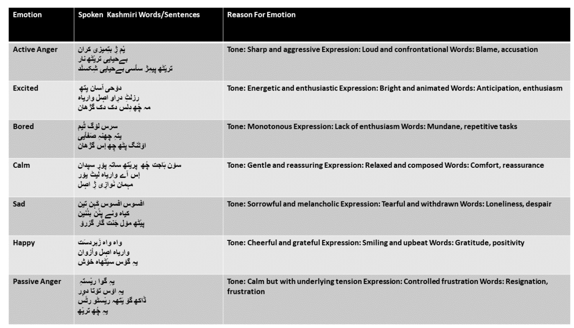
3.5. **Consent and Ethical Compliance.** A comprehensive ethics protocol governed all stages of data collection. Participants provided written and verbal informed consent, with full transparency about data usage and their right to withdraw. All personally identifiable information was anonymized, and sensitive content was excluded. Publicly available data was screened for ethical suitability and used only when permissible. Participants were also informed that the dataset may be shared for research purposes to promote open science.



(a) Overview of the human evaluation process for Kashmiri emotional speech corpus.



(b) Distribution of emotions in the annotated Kashmiri speech dataset.



Emotion	Spoken Kashmiri Words/Sentences	Reason For Emotion
Active Anger	یِم ژٕ بتِمیزی کران بےحیایی ترێٹھ نار ترێٹھ پیمِژ سأسی بےحیایی شِکسلٕد	Tone: Sharp and aggressive Expression: Loud and confrontational Words: Blame, accusation
Excited	دۆحی آسان یِتھ رزلٹِ دراو اصِل واریاہ مہ چُھ دِلس دک دک گژھان	Tone: Energetic and enthusiastic Expression: Bright and animated Words: Anticipation, enthusiasm
Bored	سرس لۆگ ٹُیم یتہِ چھنہِ صَفأیی اوٹنگ پٹھ چھ اِس گژھان	Tone: Monotonous Expression: Lack of enthusiasm Words: Mundane, repetitive tasks
Calm	سۆن بَاجت چُھ پریٚتھ ساتہِ پۆرِ سپدان اِس آے واریاہ لیٹ پۆر مہمان نَوازِی ژٕ اصِل	Tone: Gentle and reassuring Expression: Relaxed and composed Words: Comfort, reassurance
Sad	افسوس افسوس کٕہن تین کیاہ ونے پٔنُ بنُنین پیٚٹھ مۆل جَنت گار گژرو	Tone: Sorrowful and melancholic Expression: Tearful and withdrawn Words: Loneliness, despair
Happy	واہ واہ زَبردسَت واریاہ اصِل وأزوان یِہ گۆس سیٚٹھاہ خۆش	Tone: Cheerful and grateful Expression: Smiling and upbeat Words: Gratitude, positivity
Passive Anger	یِہ گوا ریٚستہِ یِہ اوس تۆتا دۆر ڈاکھ گۆ یِتھ ریٚستۆ رٹس یِہ چُھ تریٚھ	Tone: Calm but with underlying tension Expression: Controlled frustration Words: Resignation, frustration

(c) Sample Kashmiri sentences labeled with different emotions and annotated prosodic cues.

FIGURE 2. Overview of the Kashmiri emotional speech dataset creation and annotation process, including human evaluation (a), emotion distribution (b), and example annotated sentences (c).

3.6. **Kashmiri Dataset Characteristics.** Table 5 summarizes the technical and demographic characteristics of the Kashmiri emotional speech dataset.

3.7. **Auxiliary Data: Urdu, Persian, and English.** To mitigate Kashmiri's data limitations, auxiliary corpora were integrated from Urdu, Persian, and English. The choice of these languages was both linguistic and experimental: Urdu is phonologically and prosodically close to Kashmiri, Persian offers a contrast with different phonotactic structures, and English provides large, well-annotated datasets (e.g., RAVDESS, EMO-DB) as benchmarks.

TABLE 5. Summary of the curated Kashmiri speech emotion dataset

Attribute	Value
Total Speakers	166
Male/Female Ratio	88 / 78
Total Duration	5 h 30 min
Sampling Rate	16 kHz
Bit Depth	16-bit
File Format	WAV
Number of Emotion Categories	7
Labeling Method	4 Human Raters + Expert Adjudication

All auxiliary datasets were preprocessed to align with the Kashmiri pipeline (downsampling, mono conversion, normalization, silence trimming), ensuring comparability across languages. This cross-lingual setup allowed us to test both direct transfer (zero-shot) and adaptation scenarios, probing the generalization capacity of Kashmiri-trained models.

By systematically combining Kashmiri with Urdu, Persian, and English resources, this methodology not only establishes the first SER baseline for Kashmiri but also provides a reproducible framework for advancing cross-lingual SER in other underrepresented languages.

3.8. **Feature Extraction: MFCC and Spectrogram Values.** Following the creation and preprocessing of the Kashmiri emotional speech corpus, the next critical step in the methodology is acoustic feature extraction. Since emotions are primarily encoded in both the spectral and temporal dynamics of speech, we extract two complementary sets of features: Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram values. Together, these features capture short-term spectral envelopes, prosodic variations, and time-frequency energy distributions, enabling the model to recognize subtle emotional nuances. Importantly, the same pipeline was consistently applied to all auxiliary datasets (Urdu, Persian, English), ensuring uniformity and comparability across languages in subsequent cross-lingual experiments.

3.8.1. *Preprocessing and Energy Normalization.* To standardize acoustic conditions across corpora, all recordings were resampled to a fixed rate of $f_s = 16$ kHz and normalized in terms of signal energy. Variability in recording amplitude can distort feature distributions; therefore, each signal $x[n]$ was scaled to unit energy:

$$E = \sum_{n=0}^{N-1} x[n]^2, \quad x_{\text{norm}}[n] = \frac{x[n]}{\sqrt{E}} \tag{3.1}$$

Here, E denotes the total energy of the signal, and $x_{\text{norm}}[n]$ is the normalized signal. This preprocessing ensures that subsequent features depend solely on the phonetic and prosodic structure of speech rather than recording conditions or loudness variations.

3.8.2. *Spectrogram Feature Extraction.* The spectrogram provides a two-dimensional view of speech, representing the evolution of spectral energy across time. It is derived using the Short-Time Fourier Transform (STFT) applied to windowed signal frames:

$$X(k,m) = \sum_{n=0}^{L-1} x_{\text{frame}}[n] \cdot w[n] \cdot e^{-j2\pi kn/L} \tag{3.2}$$

where $x_{\text{frame}}[n]$ is the signal frame of length L, $w[n]$ is a Hamming window, k is the frequency bin index, and m the frame index.

The spectrogram magnitude is then expressed in decibel scale to better approximate human auditory perception:

$$S(k,m) = 20 \cdot \log_{10}(|X(k,m)|) \tag{3.3}$$

This logarithmic scaling emphasizes perceptually relevant variations in intensity, particularly those associated with emotional cues such as raised pitch in anger or reduced energy in sadness.

3.8.3. *MFCC Feature Extraction.* While the spectrogram captures broad spectral patterns, MFCCs model the vocal tract's spectral envelope, which is strongly shaped by emotional states. The extraction process consists of the following steps:

1. Framing: The normalized signal is segmented into overlapping frames:

$$x_{\text{frame}}[n] = x_{\text{norm}}[n + m(L-O)], \quad m = 0,1,\ldots,M-1 \tag{3.4}$$

2. Windowing: Each frame is multiplied by a Hamming window:

$$x_{\text{win}}[n] = x_{\text{frame}}[n] \cdot w[n], \quad w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) \tag{3.5}$$

3. Fourier Transform: The discrete Fourier transform (DFT) is computed:

$$X[k] = \sum_{n=0}^{L-1} x_{\text{win}}[n] \cdot e^{-j2\pi kn/L} \tag{3.6}$$

4. Mel Filter Bank: The magnitude spectrum is mapped onto the Mel scale:

$$E_m = \sum_{k=k_1}^{k_2} |X[k]|^2 H_m[k], \quad m = 1,2,\ldots,M \tag{3.7}$$

with the Mel scale defined as:

$$f_{\text{Mel}} = 2595 \cdot \log_{10}\left(1 + \frac{f_{\text{Hz}}}{700}\right) \tag{3.8}$$

5. Logarithmic Compression:

$$\text{logE}_m = \log(E_m) \tag{3.9}$$

6. Discrete Cosine Transform (DCT):

$$\text{MFCC}_c = \sum_{m=1}^{M} \text{logE}_m \cdot \cos\left[\frac{\pi c}{M}(m-0.5)\right], \quad c = 1,2,\ldots,C \tag{3.10}$$

This final step yields a compact set of cepstral coefficients capturing the smoothed spectral envelope, which is highly sensitive to prosodic shifts associated with emotion.

3.8.4. *Feature Concatenation and Alignment.* The extracted features are concatenated frame-by-frame:

$$F[n] = [S[n], \text{MFCC}[n]] \tag{3.11}$$

To maintain temporal coherence, frames are aligned and truncated to the shortest sequence length. This combined representation integrates both detailed frequency information and compact cepstral descriptors.

3.8.5. *Cross-Lingual Consistency and Dataset Balancing.* To ensure comparability across languages, the identical feature extraction pipeline was applied to Urdu, Persian, and English corpora. Furthermore, to mitigate the effects of class imbalance, equal representation of each emotion class was enforced:

$$N_{\text{balanced}} = \min\left(N_{\text{class1}}, N_{\text{class2}}, \ldots, N_{\text{classE}}\right) \tag{3.12}$$

This step ensures that the model is trained on uniformly distributed classes, preventing dominance by majority emotions and promoting fairer cross-lingual evaluation.

By uniting MFCC and spectrogram features under a standardized preprocessing framework, this study provides a feature representation that captures both fine-grained and broad emotional cues. This dual representation is especially critical for Kashmiri and other low-resource tonal languages, where subtle prosodic variations carry significant emotional meaning.

3.9. **Proposed Model.** Building upon the feature extraction process described in the previous section, the next step is to design an architecture capable of effectively modeling both the spectral-temporal characteristics of speech and the emotionally salient regions embedded within it. For this purpose, we propose a Bidirectional Long Short-Term Memory (Bi-LSTM) network augmented with an attention mechanism. This combination is particularly well-suited to Speech Emotion Recognition (SER) in Kashmiri, which is marked by tonal variations, complex prosody, and subtle intonation cues that pose challenges for conventional models. The fully connected architecture of the proposed framework is illustrated in Figure 3.

The Bi-LSTM forms the core of the model, consisting of two stacked bidirectional recurrent layers. The first layer comprises 256 hidden units, while the second layer contains 128 units. Unlike unidirectional RNNs, Bi-LSTMs process the sequence in both forward and backward directions, enabling the encoding of contextual dependencies across the entire utterance. Formally, the hidden states at time step t for forward and backward passes are denoted by $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$, respectively, and concatenated as:

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t].$$

This representation allows the network to jointly consider past and future information, which is essential for capturing the temporal evolution of emotions. ReLU activations are applied to hidden state transformations, while sigmoid activations govern the gating functions of the LSTM units.

Together, these ensure effective learning of long-range dependencies while mitigating vanishing gradient issues. To further enhance generalization and reduce overfitting, dropout layers with a rate of 0.3 are introduced after each Bi-LSTM layer, randomly deactivating neurons during training.

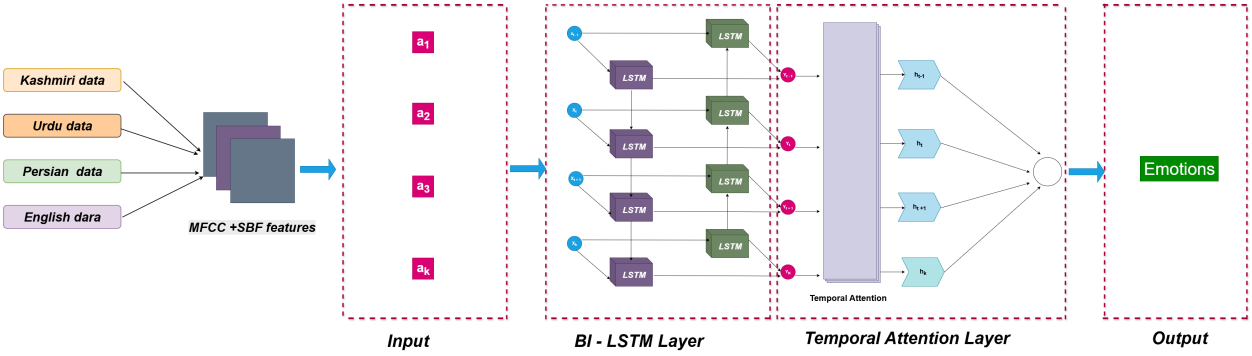


FIGURE 3. Architecture of the proposed Bi-LSTM with attention framework for multilingual speech emotion recognition.

To refine the temporal features learned by the Bi-LSTM layers, an attention mechanism is applied. This mechanism adaptively assigns importance weights α_t to each time step, enabling the model to concentrate on emotionally informative regions while down-weighting irrelevant or redundant portions of the sequence. The attention scores are computed as:

$$e_t = \tanh(W_a h_t + b_a), \quad \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)},$$

where W_a and b_a are trainable parameters, and T is the sequence length. The weighted context vector c, which summarizes the sequence based on learned attention weights, is given by:

$$c = \sum_{t=1}^{T} \alpha_t h_t.$$

This context vector enhances the model's ability to capture subtle but crucial emotional cues by focusing computational resources on the most salient frames.

The context representation is then passed through fully connected dense layers with ReLU activation, which capture higher-order feature interactions and project the temporal features into a discriminative representation space. Finally, a softmax output layer produces the probability distribution over all C emotion classes:

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)},$$

where z_i represents the logit corresponding to class i. Model training is performed using the Adam optimizer with a learning rate of 0.001, ensuring efficient gradient updates and stable convergence. The optimization objective is the sparse categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i),$$

where y_i denotes the ground-truth label for class i.

By combining the strengths of Bi-LSTMs for bidirectional temporal modeling and attention mechanisms for focusing on salient emotional cues, the proposed model provides a powerful and flexible framework for multilingual SER. This design is particularly advantageous for low-resource and tonal languages like Kashmiri, where subtle acoustic variations play a critical role in distinguishing emotions.

4. Experimental Setup

The experimental design was structured to systematically evaluate multiple deep learning architectures for Speech Emotion Recognition (SER) across Kashmiri and auxiliary languages. We explored Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory networks (Bi-LSTM), hybrid CNN–LSTM models, and Gated Recurrent Units (GRU), ensuring a comprehensive comparison of sequence modeling and feature extraction approaches. For each architecture, critical hyperparameters—including the number of hidden units, recurrent depth, dropout rate, and learning rate—were optimized through grid search to achieve the best possible performance.

While baseline models such as CNNs and GRUs achieved moderate accuracy, their limitations became evident in modeling the intricate emotional dynamics of speech. CNNs and Temporal Convolutional Networks (TCNs) effectively captured local spectral patterns but failed to account for long-range temporal dependencies. GRUs and unidirectional LSTMs learned sequential information but struggled with bidirectional contextual modeling, which is essential in tonal and low-resource languages like Kashmiri. In contrast, the Bi-LSTM augmented with attention consistently outperformed all other architectures. Its strength lies in encoding past and future temporal dependencies while adaptively focusing on emotionally salient regions of the speech signal. Based on these consistent observations, all subsequent experiments were carried out exclusively using the Bi-LSTM with attention framework to ensure methodological consistency and to fully exploit its superior representational capacity.

4.1. **Feature Inputs.**

Acoustic features served as the primary input for the model. As described in Section 3, two complementary representations were extracted: Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms. MFCCs provide compact representations of the spectral envelope and capture phonetic and prosodic patterns strongly tied to emotion. Spectrograms, on the other hand, retain detailed time–frequency dynamics, highlighting energy shifts and pitch modulations that often characterize emotional states. By combining these two feature types, the model benefits from both fine-grained phonetic cues and global prosodic contours, enabling more robust emotion recognition across languages.

4.2. **Training and Evaluation.** The Bi-LSTM with attention network was trained using the Adam optimizer with a fixed learning rate of 0.001, chosen for its adaptive gradient updates and stable convergence. Training was guided by the sparse categorical cross-entropy loss function, which is well-suited for multi-class classification with mutually exclusive emotion categories.

To ensure robustness and generalizability, a 10-fold cross-validation protocol was adopted for every dataset. In each iteration, nine folds were used for training and one for validation, allowing all samples to contribute to both phases. To prevent overfitting, early stopping with a patience of three epochs was employed, halting training once validation loss stopped improving. In parallel, model checkpointing preserved the best-performing weights, ensuring that evaluation was always performed on the most optimal model state.

Performance was evaluated across multiple dimensions. Overall classification accuracy provided a global measure of recognition capability. In addition, class-wise precision, recall, and F1-score were computed to assess the model's discrimination ability, particularly in datasets with imbalanced emotion distributions. Confusion matrices were further analyzed to identify systematic misclassifications, providing insight into emotion-specific errors and inter-class overlaps.

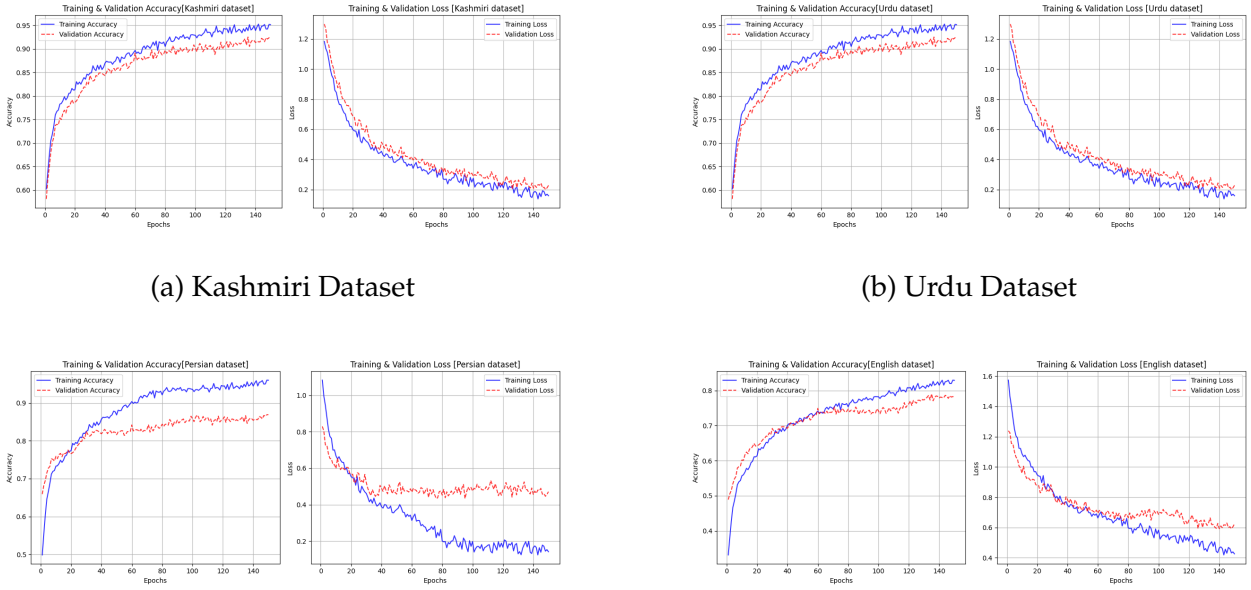
This experimental setup ensured that the evaluation of Kashmiri and auxiliary languages was both rigorous and reproducible, forming a solid foundation for the three key experiments within language evaluation, cross-lingual transfer, and progressive adaptation—presented in the subsequent sections.

4.3. **Experimental Design.** The experiments were organized into a structured series to systematically investigate factors influencing emotion recognition performance:

4.3.1. *Experiment 1: Individual Language Evaluation.* The first experiment was carried out to establish a reliable baseline for within-language emotion recognition performance. Before exploring cross-lingual transfer and adaptation, it was essential to determine how effectively the proposed Bi-LSTM with attention framework could learn and classify emotions when trained and tested on a single language at a time. This step was necessary for two reasons: first, to verify that the model can capture the intrinsic emotional dynamics of each language without external influences; and second, to provide a benchmark against which subsequent cross-lingual experiments could be objectively compared. Without such a baseline, it would be difficult to assess whether performance changes in later experiments were due to cross-lingual adaptation or limitations of the model itself.

For each dataset Kashmiri, Urdu, Persian, and English comprehensive hyperparameter tuning was performed. The number of Bi-LSTM units, recurrent depth, dropout ratios, and learning rate were optimized using a grid search strategy to balance classification accuracy and model generalization. A ten-fold cross-validation protocol was adopted to guarantee robust evaluation, ensuring that every sample contributed once to validation and nine times to training. Early stopping with a patience of three epochs was applied to halt training when validation loss

stagnated, and model checkpointing was employed to preserve the best-performing weights. To validate training reliability, learning curves of training and validation accuracy/loss were plotted for each dataset (Figures 4). These plots confirmed smooth convergence and stable learning behavior, indicating that the model was neither underfitting nor overfitting.



(a) Kashmiri Dataset

(b) Urdu Dataset

(c) Persian Dataset

(d) English Dataset

FIGURE 4. Learning behavior of the proposed Bi-LSTM with attention model across four language datasets. Each figure presents training and validation accuracy/loss curves, highlighting convergence patterns and stability for Kashmiri, Urdu, Persian, and English datasets.

The results of this experiment confirmed that the Bi-LSTM with attention network is capable of effectively modeling emotional dynamics within individual languages. For Kashmiri, which remains an underexplored language in SER research, the model achieved an accuracy of 93.2%, thereby validating the discriminative power of the extracted MFCC and spectrogram features. For Urdu, a closely related Indo-Aryan language with rich phonetic structure, performance reached 97%, highlighting the model's strong adaptability. Persian and English, which are more distant in their phonological and prosodic systems, achieved 85% and 80.05% accuracies, respectively. These results not only validate the model across typologically diverse languages but also provide evidence that tonal and prosodic similarities (e.g., between Kashmiri and Urdu) facilitate stronger recognition accuracy. The classification report for Kashmiri is presented in Table 6, while a dataset-level summary of samples, emotion categories, and final accuracies across all four languages is shown in Table 7.

In summary, Experiment 1 provided the necessary justification for proceeding with cross-lingual evaluations. By demonstrating high within-language accuracies and stable training behavior

Table 6. Classification Report for Multiclass Emotion Recognition on Kashmiri Dataset

Class	Precision	Recall	F1-Score	Support
0	0.92	0.96	0.94	34846
1	0.95	0.96	0.95	34846
2	0.94	0.93	0.93	34846
3	0.92	0.91	0.91	34846
4	0.90	0.89	0.89	34846
5	0.92	0.91	0.91	34846
6	0.94	0.93	0.93	34846
Accuracy		0.93		243922
Macro Avg	0.93	0.93	0.93	243922
Weighted Avg	0.93	0.93	0.93	243922

Table 7. Dataset Overview for Multilingual Emotion Recognition in Experiment 1

Language	Samples	Emotions	Accuracy (%)
Kashmiri	1280	7 (Active anger, Bored, Happy, Sad, Excited, Calm, Passive anger)	93.02
Urdu	1078	4 (Angry, Happy, Sad, Neutral)	97.00
Persian	1200	4 (Angry, Happy, Sad, Neutral)	85.00
English (SAVEE)	2300	7 (Angry, Happy, Sad, Neutral, Fear, Surprise, Disgust)	80.05

across diverse linguistic contexts, this experiment established a strong benchmark against which the generalizability and adaptability of the proposed framework could later be tested.

4.3.2. *Experiment 2: Cross-Lingual Evaluation with Kashmiri as the Source Language.* The second experiment was conducted to investigate the cross-lingual transferability of the proposed Bi-LSTM with attention model. In this setup, the model was trained exclusively on the Kashmiri dataset and tested on Urdu, Persian, and English datasets independently. To ensure consistency across languages, the evaluation was restricted to a common subset of four emotions: *angry*, *happy*, *sad*, and *neutral*. This restriction provided a uniform label space, allowing for a controlled comparison of cross-lingual generalization. The aim of this experiment was to determine whether emotional patterns learned from Kashmiri speech could generalize to other languages with distinct phonetic and prosodic properties.

The results demonstrate a sharp decline in performance compared to within-language evaluation. As shown in Tables 8, 9, and 10, the overall accuracies reached only 28% for Urdu, 30% for English, and 33% for Persian. These values highlight the limitations of direct transfer, as the model struggled to capture language-specific variations in emotional expression. Precision,

recall, and F1-scores reveal that *neutral* and *sad* dominate predictions, while *angry* and *happy* are frequently misclassified.

The confusion matrices presented in Figure 5 further illustrate these trends. For Persian, the model showed a strong bias toward predicting *sad*, whereas in English and Urdu, most samples converged toward *neutral*. Cross-emotion overlaps, particularly between *angry* and *happy*, were also observed, indicating that emotional boundaries learned in Kashmiri speech are not directly aligned with those of other languages. These systematic misclassifications highlight both phonetic mismatches and distributional differences in emotion corpora.

Overall, this experiment establishes that Kashmiri-trained features alone are insufficient for robust cross-lingual transfer. The results strongly justify the domain adaptation strategies explored in Experiment 3, where portions of target-language data are progressively incorporated to mitigate these mismatches.

Table 8. Classification Report on Urdu Test Set (Trained on Kashmiri)

Class	Precision	Recall	F1-Score	Support
Angry	0.28	0.22	0.25	19939
Happy	0.24	0.20	0.22	19939
Neutral	0.32	0.65	0.43	19939
Sad	0.16	0.06	0.08	19939
Accuracy		0.28		79756

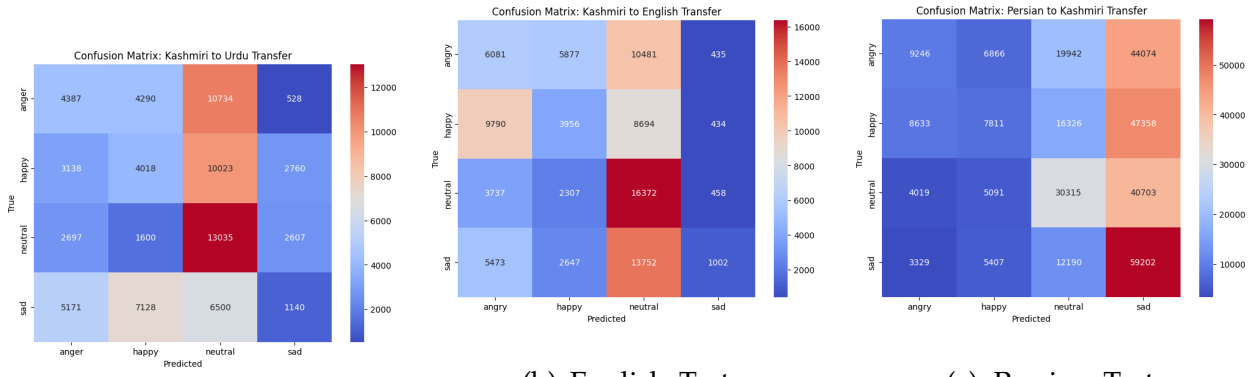
Table 9. Classification Report on English Test Set (Trained on Kashmiri)

Class	Precision	Recall	F1-Score	Support
Angry	0.24	0.27	0.25	22874
Happy	0.27	0.17	0.21	22874
Neutral	0.33	0.72	0.45	22874
Sad	0.43	0.04	0.08	22874
Accuracy		0.30		91496

4.3.3. *Experiment 3: Progressive Domain Adaptation for Cross-Lingual SER.* The cross-lingual evaluation conducted in Experiment 2 revealed a significant decline in recognition accuracy when a model trained solely on Kashmiri was applied directly to Urdu, Persian, and English datasets. This performance drop highlighted the challenges of mismatched phonetic, prosodic,

TABLE 10. Classification Report on Persian Test Set (Trained on Kashmiri)

Class	Precision	Recall	F1-Score	Support
Angry	0.37	0.12	0.18	80128
Happy	0.31	0.10	0.15	80128
Neutral	0.38	0.38	0.38	80128
Sad	0.31	0.74	0.44	80128
Accuracy		0.33		320512



(a) Urdu Test Set

(b) English Test Set

(c) Persian Test Set

FIGURE 5. Confusion matrices for cross-lingual transfer (trained on Kashmiri, tested on Urdu, English, and Persian).

and emotional distributions across languages, and underscored the need for an effective adaptation strategy. To address this limitation, we designed a progressive domain adaptation experiment that incrementally incorporated small amounts of target-language data into the Kashmiri training set. The aim was to improve the model's ability to generalize while minimizing reliance on large-scale labeled resources in the target domain.

The proposed strategy involved introducing 10%, 20%, and 30% of the target-language training samples into the original Kashmiri corpus, followed by testing the adapted model on the full target-language test set. This gradual inclusion enabled the model to capture language-specific emotional characteristics while retaining its baseline Kashmiri-learned representations. The approach was motivated by two considerations: (i) practical feasibility, as acquiring even a small portion of labeled data from the target language is often more realistic than curating a full-scale corpus, and (ii) efficiency, since exposing the model to incremental target-language data can reveal how quickly performance improves with minimal adaptation effort.

The results demonstrated a clear and consistent trend of performance improvement across all languages. For English, accuracy increased from 34% without adaptation to 61%, 78%, and 83% at the 10%, 20%, and 30% adaptation levels, respectively. A similar trend was observed for Persian, where accuracy rose from 33% to 59%, 71%, and 81%. Urdu exhibited the most substantial improvement, with accuracy climbing from 25% to 80%, 87%, and 89%. These outcomes validate the effectiveness of progressive adaptation and confirm that even limited exposure to target-language data is sufficient to substantially boost cross-lingual performance. The especially strong gains for Urdu can be attributed to its close linguistic and prosodic affinity with Kashmiri, while the comparatively lower but still significant improvements for English and Persian reflect their greater phonological distance.

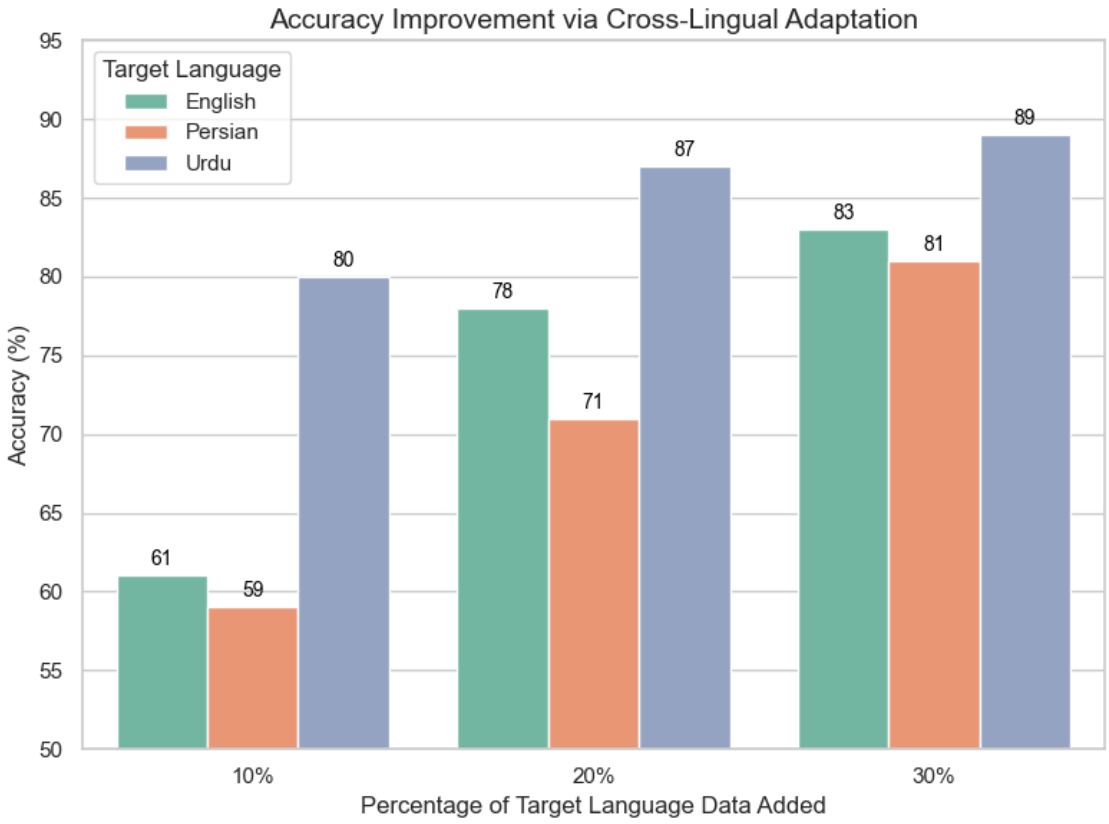


FIGURE 6. Progressive domain adaptation results: accuracy improvements for English, Persian, and Urdu as incremental portions (10%, 20%, and 30%) of target language data are incorporated into the Kashmiri-trained model.

The progressive adaptation results carry two important implications. First, they demonstrate that robust cross-lingual emotion recognition is feasible even in low-resource settings, provided that minimal supervised target-language data is available. Second, the steep improvement between 10% and 20% adaptation levels suggests that substantial performance gains can be achieved with relatively low annotation costs. To illustrate these findings, Figure 6 presents a

bar chart comparing baseline performance with progressively adapted results across the three target languages, highlighting the consistent upward trajectory. Together, these findings establish progressive domain adaptation as an effective strategy for bridging cross-lingual gaps in SER, particularly for under-resourced languages such as Kashmiri.

5. Results and Discussion

The structured experiments conducted in this work provide a comprehensive view of how the proposed Bi-LSTM with attention framework performs under within-language, zero-shot transfer, and progressive adaptation settings. Instead of focusing solely on raw numerical outcomes, this section highlights key patterns, interprets underlying causes, and connects them to broader challenges in multilingual speech emotion recognition.

When trained and evaluated on the same language, the model consistently demonstrated robust performance, achieving accuracies above 90% for Kashmiri and Urdu, and competitive results for Persian and English. These outcomes highlight the capacity of the proposed architecture to capture emotion-specific dynamics effectively when phonetic and prosodic cues are consistent across training and testing. However, the picture changes significantly in zero-shot transfer experiments, where the model was trained on Kashmiri and directly tested on Urdu, Persian, and English. Here, performance dropped sharply to the 25–34% range, exposing the difficulty of transferring emotional representations across languages without adaptation. The confusion matrices shed light on this decline: predictions on Persian were dominated by the *sad* class, while in Urdu and English the model tended to default heavily to *neutral*. These systematic biases are consistent with the t-SNE visualizations as illustrated in figure 8, which reveal heavy overlap among high-arousal emotions such as *angry* and *happy*, particularly in Kashmiri and Persian, where emotional boundaries are less distinct in the feature space.

The adaptation experiments provide a more optimistic perspective. By incorporating even a small portion of target-language data into Kashmiri training, accuracy improved dramatically across all target languages. For instance, the model's performance on Urdu rose from 25% in zero-shot settings to nearly 90% with just 30% of the target data, while Persian and English also experienced gains exceeding 80% as summarized in Figure 7. These results underscore the value of progressive adaptation: a modest investment in target language annotation yields substantial improvements, effectively bridging the gap between zero shot transfer and within language performance. The most pronounced benefits were observed for Urdu, which can be attributed to its close phonological and prosodic affinity with Kashmiri, while Persian and English showed weaker but still notable improvements. The summary bar chart across all three experiments illustrates this trajectory, highlighting how adaptation transforms cross-lingual SER from limited to highly effective with relatively low additional data requirements.

A central observation is the strong contrast between robust within-language performance and the sharp decline in zero-shot transfer. While the model learned emotion-specific patterns

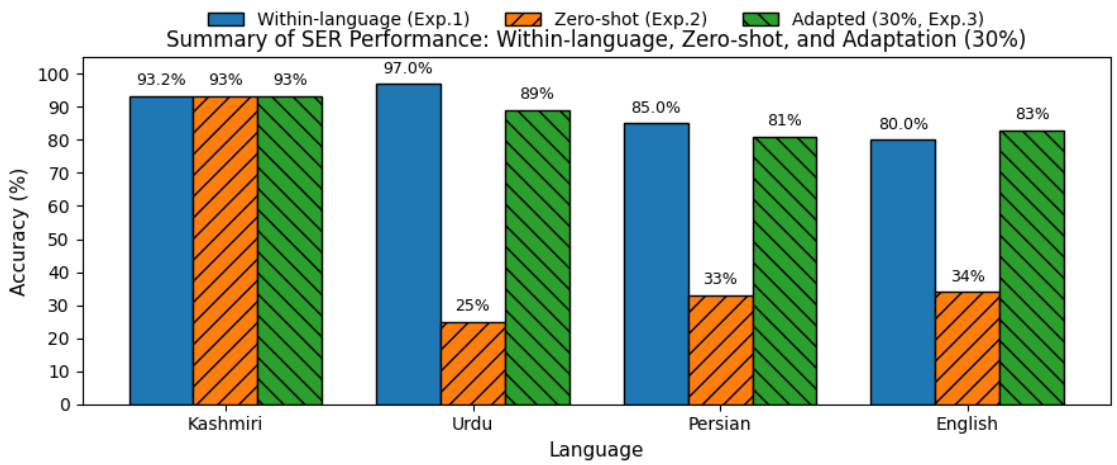


FIGURE 7. Summary of Speech Emotion Recognition (SER) performance across four languages (Kashmiri, Urdu, Persian, English). Results are reported for three experimental setups: (1) within-language training and testing, (2) zero-shot transfer from Kashmiri to the target language, and (3) adaptation using 30% target-language data.

effectively when trained and tested on the same language, its direct transfer across languages proved less reliable. Confusion matrices revealed systematic biases: in Persian, a disproportionate number of predictions gravitated toward the sad class, while in Urdu and English, the model frequently defaulted to neutral. Rather than a shortcoming of the framework, these outcomes reflect two deeper challenges in multilingual SER: (i) cross-linguistic differences in prosodic and phonetic expression of emotions, and (ii) natural variation in class distributions across datasets. For instance, Persian data contained a higher prevalence of sad utterances, while Urdu and English exhibited a more balanced spread. Such differences naturally bias model predictions but, importantly, they also shed light on how emotional cues are encoded differently across linguistic and cultural contexts.

The progressive adaptation experiments demonstrate that these barriers are not fundamental. Even modest inclusion of target-language samples in training rapidly improved performance across all datasets, with Urdu showing the steepest gains due to its close linguistic proximity to Kashmiri. These findings confirm that distributional mismatches and cross-lingual prosodic shifts can be effectively mitigated by adaptation, highlighting the feasibility of building scalable SER systems for low-resource languages with limited annotation costs.

Another key finding concerns the relative transferability of emotions. The sad emotion consistently emerged as the most stable across languages, forming clearer clusters in t-SNE space and achieving higher recognition rates in cross-lingual evaluations.

This robustness likely stems from its distinct acoustic correlates—lower pitch, slower rate, and reduced energy that remain consistent across languages. By contrast, high-arousal emotions such

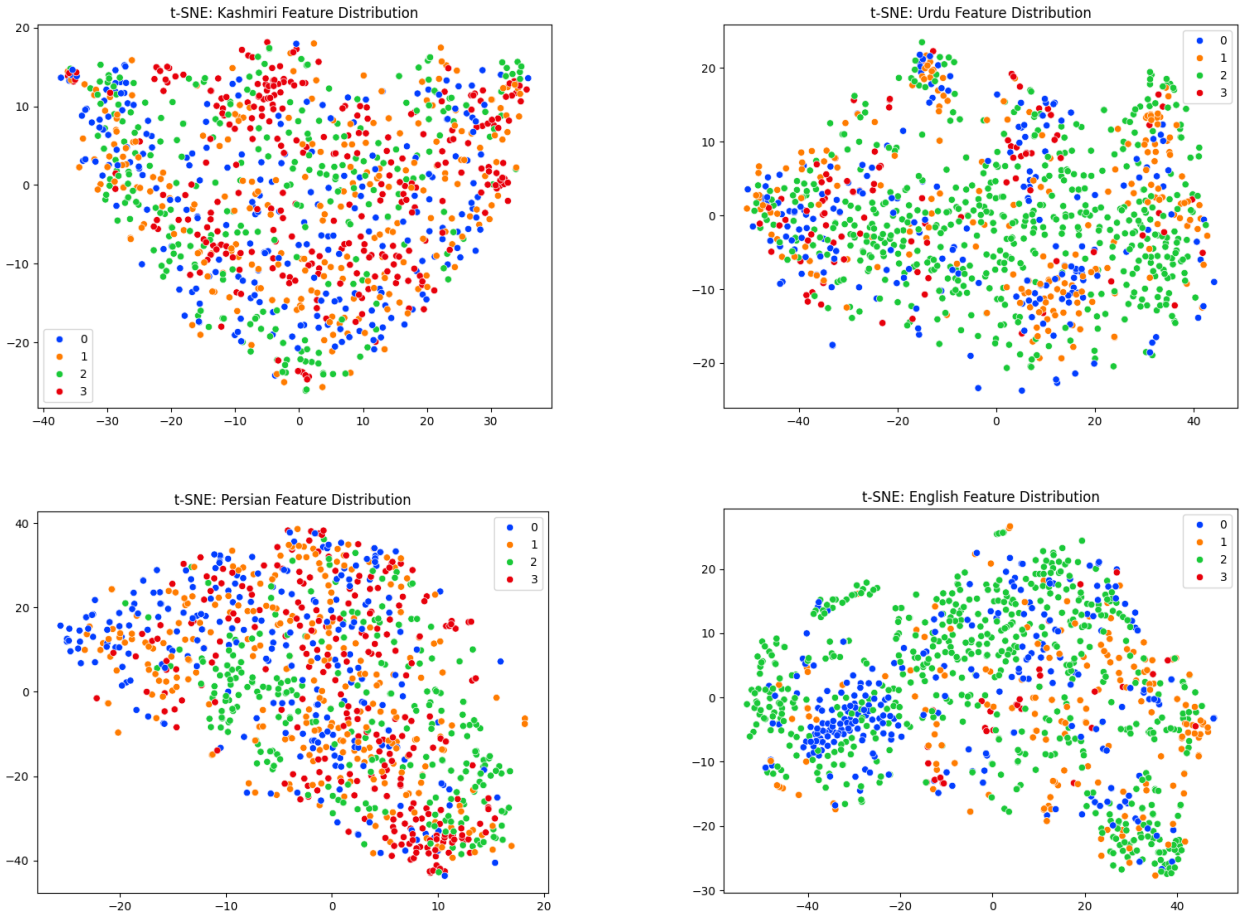


FIGURE 8. t-SNE visualizations of feature distributions across the four datasets. English and Urdu show clearer separations, indicating higher discriminability, while Kashmiri and Persian exhibit substantial overlaps among classes.

as angry and happy overlapped heavily with neutral, reflecting their reliance on prosodic cues that are more language-specific. This insight underscores the importance of designing feature representations that can better disentangle high-arousal categories in multilingual contexts.

In summary, the experiments establish three key insights: (i) the Bi-LSTM with attention framework provides strong baselines for Kashmiri and auxiliary languages but requires adaptation for reliable cross-lingual transfer; (ii) observed prediction biases arise not from methodological shortcomings but from intrinsic linguistic and distributional differences across datasets; and (iii) progressive domain adaptation offers a practical, data-efficient pathway to mitigate these effects and achieve competitive performance across diverse languages. Together with classification reports, confusion matrices, t-SNE plots, and bar chart summaries, these findings provide a multi-layered picture of system behavior and highlight the promise of cross-lingual approaches for extending SER to low-resource languages.

6. Conclusion and Future Directions

This study marks an important step toward advancing Speech Emotion Recognition (SER) for Kashmiri, an underexplored and low-resource language. While our framework demonstrated competitive performance within-language and revealed valuable insights into the challenges of cross-lingual transfer, it also opens several promising avenues for future exploration that extend beyond the current state of research.

One immediate direction lies in the integration of advanced cross-lingual adaptation strategies. While progressive domain adaptation with limited target-language data proved effective in this work, more sophisticated methods such as adversarial feature alignment, meta-learning, and self-supervised representation learning offer the potential to further reduce the gap between languages. Recent multilingual pretraining paradigms, such as wav2vec 2.0, HuBERT, and Whisper, which capture universal acoustic–prosodic features, could provide powerful initialization for SER models, thereby enabling robust generalization even in extremely low-resource settings. Incorporating these paradigms into Kashmiri SER represents a novel research direction that could set benchmarks for other endangered and resource-scarce languages.

Another compelling extension involves the transition from unimodal to multimodal emotion recognition. Human emotion is rarely expressed through speech alone; it is embedded in facial expressions, gestures, physiological signals, and linguistic context. By combining acoustic cues with visual or textual modalities, future work can address ambiguities observed in this study—for instance, the overlap between anger, happiness, and neutrality in the acoustic space. Multimodal fusion frameworks, especially those employing attention-based cross-modal alignment, may therefore substantially improve both accuracy and interpretability.

Equally important is the expansion of emotional corpora. The Kashmiri dataset curated in this work provides an essential foundation, but future efforts should aim for greater scale and diversity, incorporating spontaneous and conversational speech, multiple dialects, and a larger speaker pool. Parallel initiatives in other South Asian languages, such as Sindhi, Pashto, or Konkani, would allow the creation of a multilingual benchmark for the region, supporting comparative studies and paving the way for multilingual SER models capable of cross-lingual generalization at scale.

Finally, beyond methodological advancements, future research should explore real-world applications where robust emotion recognition in Kashmiri and related languages could have tangible impact. Examples include mental health monitoring systems, affect-aware virtual assistants, and multilingual call-center analytics, where emotion-aware interfaces can enhance accessibility, inclusivity, and user experience. Deploying SER in these domains would not only validate its technological promise but also demonstrate the societal value of preserving and digitizing the emotional richness of low-resource languages.

In summary, future directions for this line of research converge on three key pillars: (i) advancing cross-lingual adaptation through self-supervised and transfer learning approaches, (ii) expanding into multimodal emotion recognition to capture the holistic nature of human affect, and (iii) scaling

and diversifying low-resource corpora for wider linguistic inclusivity. Together, these avenues chart a novel and impactful research trajectory, ensuring that the work initiated here evolves into a broader scientific and societal contribution.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] S. Madanian, T. Chen, O. Adeleye, J.M. Templeton, C. Poellabauer, et al., Speech Emotion Recognition Using Machine Learning — A Systematic Review, Intell. Syst. Appl. 20 (2023), 200266. https://doi.org/10.1016/j.iswa.2023.200266.

[2] G.H. Mohmad Dar, R. Delhibabu, Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review, IEEE Access 12 (2024), 151122–151152. https://doi.org/10.1109/access.2024.3476960.

[3] G. Alhussein, I. Ziogas, S. Saleem, L.J. Hadjileontiadis, Speech Emotion Recognition in Conversations Using Artificial Intelligence: A Systematic Review and Meta-Analysis, Artif. Intell. Rev. 58 (2025), 198. https://doi.org/10.1007/s10462-025-11197-8.

[4] R. Zhao, X. Jiang, F. Richard Yu, V.C.M. Leung, T. Wang, et al., Leveraging Cross-Attention Transformer and Multifeature Fusion for Cross-Linguistic Speech Emotion Recognition, IEEE Internet Things J. 12 (2025), 50653–50664. https://doi.org/10.1109/jiot.2025.3613687.

[5] S. Zhang, R. Liu, X. Tao, X. Zhao, Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives, Front. Neurorobotics 15 (2021), 784514. https://doi.org/10.3389/fnbot.2021.784514.

[6] R. Ullah, M. Asif, W.A. Shah, F. Anjam, I. Ullah, et al., Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer, Sensors 23 (2023), 6212. https://doi.org/10.3390/s23136212.

[7] N. Saleem, J. Gao, R. Irfan, A. Almadhor, H.T. Rauf, et al., DeepCNN: Spectro-Temporal Feature Representation for Speech Emotion Recognition, CAAI Trans. Intell. Technol. 8 (2023), 401–417. https://doi.org/10.1049/cit2.12233.

[8] G.M. Dar, R. Delhibabu, Exploring Emotion Detection in Kashmiri Audio Reviews Using the Fusion Model of CNN, LSTM, and RNN: Gender-Specific Speech Patterns and Performance Analysis, Int. J. Inf. Technol. (2024). https://doi.org/10.1007/s41870-024-02105-4.

[9] C. Barhoumi, Y. BenAyed, Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation, Artif. Intell. Rev. 58 (2024), 49. https://doi.org/10.1007/s10462-024-11065-x.

[10] E.A. Alkhamali, A. Allinjawi, R.B. Ashari, Combining Transformer, Convolutional Neural Network, and Long Short-Term Memory Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition

in Distance Education Classrooms, Appl. Sci. 14 (2024), 5050. https://doi.org/10.3390/app14125050.

[11] V. Bhardwaj, An Experimental Framework of Speaker Independent Speech Recognition System for Kashmiri Language (k-Asr) System Using Sphinx, Int. J. Emerg. Trends Sci. Technol. 04 (2017), 5348–5352. https://doi.org/10.18535/ijetst/v4i7.07.

[12] G.M. Dar, R. Delhibabu, Emotion Recognition in Kashmiri Speech: Evaluating Coefficient-Based Acoustic Features Using Bidirectional LSTM Networks, Procedia Comput. Sci. 258 (2025), 1909–1921. https://doi.org/10.1016/j.procs.2025.04.442.

[13] Y.R. Dar, A. Nazir, M. Ahmed, Acoustic Analysis of Vowels in Kashmiri-Speaking Adolescents With Down Syndrome, J. Appl. Linguist. Lang. Res. 7 (2020), 168–175.

[14] K. Scherer, Vocal Communication of Emotion: A Review of Research Paradigms, Speech Commun. 40 (2003), 227–256. https://doi.org/10.1016/s0167-6393(02)00084-5.

[15] Y. Gao, L. Wang, J. Liu, J. Dang, S. Okada, Adversarial Domain Generalized Transformer for Cross-Corpus Speech Emotion Recognition, IEEE Trans. Affect. Comput. 15 (2024), 697–708. https://doi.org/10.1109/taffc.2023.3290795.

[16] M. Agarla, S. Bianco, L. Celona, P. Napoletano, A. Petrovsky, et al., Semi-Supervised Cross-Lingual Speech Emotion Recognition, Expert Syst. Appl. 237 (2024), 121368. https://doi.org/10.1016/j.eswa.2023.121368.

[17] S.G. Koolagudi, R. Reddy, J. Yadav, K.S. Rao, IITKGP-SEHSC: Hindi Speech Corpus for Emotion Analysis, in: 2011 International Conference on Devices and Communications (ICDeCom), IEEE, 2011, pp. 1-5. https://doi.org/10.1109/icdecom.2011.5738540.

[18] A. Asghar, S. Sohaib, S. Iftikhar, M. Shafi, K. Fatima, An Urdu Speech corpus for Emotion Recognition, PeerJ Comput. Sci. 8 (2022), e954. https://doi.org/10.7717/peerj-cs.954.

[19] S. Mohanty, B.K. Swain, Emotion Recognition Using Fuzzy K-Means from Oriya Speech, Int. J. Comput. Commun. Technol. 1 (2011), 24–28. https://doi.org/10.47893/ijcct.2011.1066.

[20] A. Geethashree, D.J. Ravi, Kannada Emotional Speech Database: Design, Development and Evaluation, Lecture Notes in Networks and Systems, Vol. 14, Springer, Singapore, 2017: pp. 135–143. https://doi.org/10.1007/978-981-10-5146-3_14.

[21] A. Jacob, Modelling Speech Emotion Recognition Using Logistic Regression and Decision Trees, Int. J. Speech Technol. 20 (2017), 897–905. https://doi.org/10.1007/s10772-017-9457-6.

[22] G. Agarwal, H. Om, Performance of Deer Hunting Optimization Based Deep Learning Algorithm for Speech Emotion Recognition, Multimed. Tools Appl. 80 (2020), 9961–9992. https://doi.org/10.1007/s11042-020-10118-x.

[23] S. Sultana, M.S. Rahman, M.R. Selim, M.Z. Iqbal, SUST Bangla Emotional Speech Corpus (SUBESCO): An Audio-Only Emotional Speech Corpus for Bangla, PLOS ONE 16 (2021), e0250173. https://doi.org/10.1371/journal.pone.0250173.

[24] Z.S. Syed, S. Ali, M. Shehram, A. Shah, Introducing the Urdu-Sindhi Speech Emotion Corpus: A Novel Dataset of Speech Recordings for Emotion Recognition for Two Low-Resource

Languages, Int. J. Adv. Comput. Sci. Appl. 11 (2020), 01104104. https://doi.org/10.14569/IJACSA.2020.01104104.

[25] A.K. Samantaray, K. Mahapatra, B. Kabi, A. Routray, A Novel Approach of Speech Emotion Recognition with Prosody, Quality and Derived Features Using SVM Classifier for a Class of North-Eastern Languages, in: 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), IEEE, 2015: pp. 372–377. https://doi.org/10.1109/ReTIS.2015.7232907.

[26] S.A. Ali, A. Khan, N. Bashir, Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus, Int. J. Inf. Technol. Comput. Sci. 7 (2015), 54–59. https://doi.org/10.5815/ijitcs.2015.02.07.

[27] R.V. Darekar, A.P. Dhande, Emotion Recognition from Marathi Speech Database Using Adaptive Artificial Neural Network, Biol. Inspir. Cogn. Arch. 23 (2018), 35–42. https://doi.org/10.1016/j.bica.2018.01.002.

[28] J. Basu, S. Majumder, Performance Evaluation of Language Identification on Emotional Speech Corpus of Three Indian Languages, in: Intelligence Enabled Research. Advances in Intelligent Systems and Computing, Springer Singapore, 2020: pp. 55–63. https://doi.org/10.1007/978-981-15-9290-4_6.

[29] P. Dhar, S. Guha, A System to Predict Emotion from Bengali Speech, Int. J. Math. Sci. Comput. 7 (2021), 26–35. https://doi.org/10.5815/ijmsc.2021.01.04.

[30] B. Fernandes, K. Mannepalli, Speech Emotion Recognition Using Deep Learning LSTM for Tamil Language, Pertanika J. Sci. Technol. 29 (2021), 1915–1936. https://doi.org/10.47836/pjst.29.3.33.

[31] V.P. Tank, S.K. Hadia, Creation of Speech Corpus for Emotion Analysis in Gujarati Language and Its Evaluation by Various Speech Parameters, Int. J. Electr. Comput. Eng. (IJECE) 10 (2020), 4752. https://doi.org/10.11591/ijece.v10i5.pp4752-4758.

[32] S. Aziz, N.H. Arif, S. Ahbab, S. Ahmed, T. Ahmed, et al., Improved Speech Emotion Recognition in Bengali Language Using Deep Learning, in: 2023 26th International Conference on Computer and Information Technology (ICCIT), IEEE, 2023: pp. 1–6. https://doi.org/10.1109/ICCIT60459.2023.10441053.

[33] T.J. Sefara, The Effects of Normalisation Methods on Speech Emotion Recognition, in: 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), IEEE, 2019. https://doi.org/10.1109/IMITEC45504.2019.9015895.

[34] P. Harar, R. Burget, M.K. Dutta, Speech Emotion Recognition with Deep Learning, in: 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2017: pp. 137–140. https://doi.org/10.1109/SPIN.2017.8049931.

[35] Mustaqeem, M. Sajjad, S. Kwon, Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM, IEEE Access 8 (2020), 79861–79875. https://doi.org/10.1109/ACCESS.2020.2990405.

[36] K.H. Lee, D.H. Kim, Design of a Convolutional Neural Network for Speech Emotion Recognition, in: 2020 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2020: pp. 1332–1335. https://doi.org/10.1109/ICTC49870.2020.9289227.

[37] Z. Han, J. Wang, Speech Emotion Recognition Based on Deep Learning and Kernel Nonlinear PSVM, in: 2019 Chinese Control And Decision Conference (CCDC), IEEE, 2019: pp. 1426–1430. https://doi.org/10.1109/CCDC.2019.8832414.

[38] J. Wang, Z. Han, Research on Speech Emotion Recognition Technology Based on Deep and Shallow Neural Network, in: 2019 Chinese Control Conference (CCC), IEEE, 2019: pp. 3555–3558. https://doi.org/10.23919/ChiCC.2019.8866568.

[39] G. Liu, W. He, B. Jin, Feature Fusion of Speech Emotion Recognition Based on Deep Learning, in: 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), IEEE, 2018: pp. 193–197. https://doi.org/10.1109/icnidc.2018.8525706.

[40] S. Davis, P. Mermelstein, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (1980), 357–366. https://doi.org/10.1109/TASSP.1980.1163420.

[41] J. Li, L. Deng, Y. Gong, R. Haeb-Umbach, An Overview of Noise-Robust Automatic Speech Recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (2014), 745–777. https://doi.org/10.1109/TASLP.2014.2304637.

[42] C. Oflazoglu, S. Yildirim, Recognizing Emotion from Turkish Speech Using Acoustic Features, EURASIP J. Audio Speech Music. Process. 2013 (2013), 26. https://doi.org/10.1186/1687-4722-2013-26.

[43] S. Agarwalla, K.K. Sarma, Machine Learning Based Sample Extraction for Automatic Speech Recognition Using Dialectal Assamese Speech, Neural Netw. 78 (2016), 97–111. https://doi.org/10.1016/j.neunet.2015.12.010.

[44] A. Mohanty, R.C. Cherukuri, Whispered Speech Emotion Recognition with Gender Detection Using BiLSTM and DCNN, J. Inf. Syst. Telecommun. 12 (2024), 152–161. https://doi.org/10.61186/jist.43703.12.46.152.

[45] J.H. Chowdhury, S. Ramanna, K. Kotecha, Speech Emotion Recognition with Light Weight Deep Neural Ensemble Model Using Hand Crafted Features, Sci. Rep. 15 (2025), 11824. https://doi.org/10.1038/s41598-025-95734-z.

[46] S. Leem, D. Fulford, J. Onnela, D. Gard, C. Busso, Selective Acoustic Feature Enhancement for Speech Emotion Recognition with Noisy Speech, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (2024), 917–929. https://doi.org/10.1109/TASLP.2023.3340603.

[47] X. Yuanchao, C. Zhiming, K. Xiaopeng, Improved Pitch Shifting Data Augmentation for Ship-Radiated Noise Classification, Appl. Acoust. 211 (2023), 109468. https://doi.org/10.1016/j.apacoust.2023.109468.

[48] K. Kaur, P. Singh, Impact of Feature Extraction and Feature Selection Algorithms on Punjabi Speech Emotion Recognition Using Convolutional Neural Network, ACM Trans. Asian Low-Resource Lang. Inf. Process. 21 (2022), 1–23. https://doi.org/10.1145/3511888.

[49] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[50] A. Graves, J. Schmidhuber, Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures, Neural Netw. 18 (2005), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

[51] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to Forget: Continual Prediction with LSTM, Neural Comput. 12 (2000), 2451–2471. https://doi.org/10.1162/089976600300015015.

[52] D. Issa, M. Fatih Demirci, A. Yazici, Speech Emotion Recognition with Deep Convolutional Neural Networks, Biomed. Signal Process. Control. 59 (2020), 101894. https://doi.org/10.1016/j.bspc.2020.101894.

[53] T. Ozseven, Evaluation of the Effect of Frame Size on Speech Emotion Recognition, in: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), IEEE, 2018: pp. 1–4. https://doi.org/10.1109/ISMSIT.2018.8567303.

[54] S. Ntalampiras, N. Fakotakis, Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition, IEEE Trans. Affect. Comput. 3 (2012), 116–125. https://doi.org/10.1109/T-AFFC.2011.31.

[55] S. Bai, J.Z. Kolter, V. Koltun, An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, arXiv:1803.01271 (2018). https://doi.org/10.48550/arXiv.1803.01271.

[56] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation, arXiv:1406.1078 (2014). https://doi.org/10.48550/ARXIV.1406.1078.

[57] Y. Kim, C. Denton, L. Hoang, A.M. Rush, Structured Attention Networks, arXiv:1702.00887 (2017). https://doi.org/10.48550/ARXIV.1702.00887.

[58] J. Shen, R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, et al., Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018: pp. 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368.

[59] B. Su, C. Chang, Y. Lin, C. Lee, Improving Speech Emotion Recognition Using Graph Attentive Bi-Directional Gated Recurrent Unit Network, in: Proceedings of INTERSPEECH 2020, pp. 506–510, 2020. https://doi.org/10.21437/Interspeech.2020-1733.

[60] K. Manohar, E. Logashanmugam, Speech-Based Human Emotion Recognition Using CNN and LSTM Model Approach, in: Smart Innovation, Systems and Technologies, Springer, Singapore, 2022: pp. 85–93. https://doi.org/10.1007/978-981-16-9669-5_8.

[61] S. Sultana, M.Z. Iqbal, M.R. Selim, M.M. Rashid, M.S. Rahman, Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks, IEEE Access 10 (2022), 564–578. https://doi.org/10.1109/ACCESS.2021.3136251.

[62] S. Wang, X. Fu, K. Ding, C. Chen, H. Chen, et al., Federated Few-Shot Learning, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2023, pp. 2374–2385. https://doi.org/10.1145/3580305.3599347.

[63] J. Yang, J. Liu, K. Huang, J. Xia, Z. Zhu, et al., Single- and Cross-Lingual Speech Emotion Recognition Based on WavLM Domain Emotion Embedding, Electronics 13 (2024), 1380. https://doi.org/10.3390/electronics13071380.

[64] S.A.M. Zaidi, S. Latif, J. Qadir, Cross-Language Speech Emotion Recognition Using Multimodal Dual Attention Transformers, arXiv:2306.13804 (2023). https://doi.org/10.48550/ARXIV.2306.13804.