Cross-Domain Evaluation of Transformer-Based Models for Punjabi Speech Emotion Recognition

Fatima Tu Zahra, Kulsoom Asim, Sandesh Kumar, Abdul Samad
Habib University, Dhanani School of Science and Engineering, Karachi, Pakistan
ft09259@st.habib.edu.pk, ka08051@st.habib.edu.pk,
sandesh.kumar@sse.habib.edu.pk, abdul.samad@sse.habib.edu.pk

Abstract

Speech Emotion Recognition (SER) is an important part of human–computer interaction, but most existing research focuses on high-resource languages, with very limited work on regional languages such as Punjabi. This paper focuses on detecting emotions from Punjabi speech using machine learning and deep learning techniques. We curated our own Punjabi speech emotion dataset using volunteer recordings and real-world sources, covering four emotion classes: angry, happy, sad, and neutral. The data was preprocessed for consistency and evaluated using a multi-strategy framework ($E1$–$E4$) to test domain generalization. Three models were evaluated: CNN-BiLSTM, ResNet-34, and the transformer-based Wav2Vec 2.0. Among these, the ResNet-34 model performed the best in the combined-domain strategy ($E4$), achieving a test accuracy of 96%. While cross-corpus evaluations ($E2$, $E3$) highlighted challenges in generalizing to neutral emotions, the model achieved perfect scores for happy and sad classes in $E4$. These results demonstrate the effectiveness of residual networks and combined-domain training for emotion recognition in low-resource languages and highlight the potential for further work on Punjabi SER.

Keywords: Speech emotion recognition, Punjabi speech, low-resource languages, deep learning, transformer models, Wav2Vec, ResNet, machine learning

1. Introduction

Punjabi is spoken by approximately 125 million people worldwide (SIL International, 2023) and is one of the major languages of South Asia. Despite its widespread use, publicly accessible resources for speech processing remain limited, particularly for Speech Emotion Recognition (SER). SER aims to identify human emotional states from vocal cues and has applications in mental health monitoring, human–computer interaction, call center analytics, and educational technologies (Ververidis & Kotropoulos, 2006; El Ayadi et al., 2011). However, most existing SER research focuses primarily on high-resource languages such as English and Mandarin, leaving low-resource languages like Punjabi comparatively underexplored.

A central challenge in low-resource SER is the scarcity of diverse and naturalistic datasets. Existing Punjabi resources, such as the RASA corpus (Varadhan et al., 2024), were originally developed for expressive text-to-speech (TTS) and consist of studio-recorded speech from only two speakers under controlled acoustic conditions. Although such datasets are valuable for synthesis tasks, they may not adequately reflect the acoustic variability present in real-world environments and may therefore limit model generalization across speakers and domains.

To address these limitations, we introduce a newly curated Punjabi SER dataset consisting of 893 labeled utterances across four emotion categories: happy, sad, angry, and neutral. The dataset combines volunteer recordings captured in both clean and noisy real-world settings with samples sourced from audiovisual media. This design increases speaker diversity and acoustic variability compared to studio-only corpora, providing a more realistic benchmark for emotion recognition in Punjabi speech.

In addition to dataset construction, we systematically investigate cross-dataset generalization in Punjabi SER. We design four experimental strategies to evaluate in-domain speaker-dependent evaluation on our curated dataset (E1), cross-domain transfer from RASA to curated speech (E2), reverse transfer from curated to RASA (E3), and combined-domain training using both datasets (E4). Through these controlled experiments, we analyze the impact of domain mismatch and speaker variation on model robustness. We further examine the effectiveness of transformer-based self-supervised speech models, particularly Wav2Vec 2.0, in both in-domain and cross-domain settings, while also comparing against ResNet-34 and CNN-BiLSTM baselines, with a focus on understanding generalization behavior under domain shift in low-resource South Asian languages.

2. Research Question

This study addresses the following research question: “Given an audio clip in Punjabi, how accurately can computational models detect the underlying emotional state across different speakers and recording conditions?”

Specifically, we evaluate model performance

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 59–67
16 May 2026.

on four emotion categories — angry, happy, sad, and neutral — under four experimental strategies: in-domain evaluation on our curated dataset (E1), cross-domain transfer from RASA to curated speech (E2), reverse transfer from curated speech to RASA (E3), and combined-domain training using both datasets (E4). This design allows us to assess both in-domain accuracy and robustness under domain and speaker variability.

3. Literature Review

Recent advances in Speech Emotion Recognition (SER) have significantly improved the ability of models to detect emotions in speech, even for low-resource languages like Punjabi. Modern deep learning approaches capture both temporal and contextual information in speech, which is crucial for modeling emotion dynamics. For example, Graph-LSTM networks treat speech frames as connected nodes, allowing models to track emotional changes across utterances (Li et al., 2023). Similarly, CNN–BiLSTM architectures, especially when combined with noise and spectral augmentations, demonstrate that combining convolutional feature extraction with sequential modeling enhances robustness under varying acoustic conditions (Barhoumi & Ben Ayed, 2023). However, most existing studies rely on controlled datasets with acted speech, limiting generalization to real-world, naturalistic recordings and low-resource languages (El Ayadi et al., 2011).

In Punjabi SER, early work by Singla & Singh (2022) introduced a semi-natural speech dataset of 985 utterances from Punjabi films and television shows, annotated for four emotions: happy, sad, angry, and neutral. They extracted prosodic, spectral, and wavelet features and tested classifiers such as Random Forest, SVM, Naïve Bayes, and Decision Trees, achieving a maximum accuracy of 61.65% with Random Forest. This study highlighted the importance of feature engineering for low-resource language SER. Later, Singla et al. (2024) applied deep learning, using CNNs on a larger dataset of 9,000 utterances from films, web series, and social media. Preprocessing, segmentation, and augmentation addressed class imbalance, and their model achieved 69% accuracy, demonstrating the advantage of high-level spectral features and deep architectures. Chaitanya Singla and Sukhdev Singh (Singla & Singh, 2022) also released PEMO, a large-scale natural Punjabi speech emotion dataset of 22,000 utterances with strict label agreement, capturing real-life emotional expressions and providing a valuable benchmark for future models. Kaur & Singh (2022) further emphasized the role of feature selection and extraction in improving Punjabi SER performance using CNNs.

Despite these contributions, critical gaps remain. Many datasets, including the 9,000-utterance corpus (Singla et al., 2024) and PEMO (Singla & Singh, 2022), are not publicly accessible, limiting reproducibility and comparative evaluation. Additionally, prior studies rarely evaluate model robustness under domain shift or cross-dataset scenarios, leaving the generalizability of models to naturalistic speech untested.

4. Dataset

4.1. Source and Collection

Our Punjabi Speech Emotion Recognition (SER) dataset draws from two complementary sources: volunteer recordings and publicly available audiovisual media. This dual-source design was adopted to maximise acoustic and speaker diversity, combining controlled scripted utterances with naturalistic expressive speech extracted from real-world content.

For the volunteer component, ten participants contributed recordings spanning four target emotion categories: anger, happiness, sadness, and neutral. Participants ranged in age from 21 to 60 years, with an equal gender split of five male and five female speakers. Seven participants were native Punjabi speakers; the remaining three were fluent non-native speakers. This demographic composition is summarised in Table 1.

Each participant was provided with five prewritten scripts, each containing ten sentences per emotion class (40 sentences per script), yielding a pool of scripted utterances. Recordings were conducted in both quiet and acoustically noisy environments to increase domain diversity. Individual utterance durations ranged from approximately 2 to 7 seconds depending on speaker fluency.

All audio files were standardised to mono 16-bit PCM WAV format at a sampling rate of 16 kHz. Approximately half of the recordings were denoised and amplitude-normalised in post-processing; the remainder were retained in their original acoustic condition to preserve variability. The neutral-category samples were sourced from the Shrutilipi ASR corpus (Bhogale et al., 2022), a publicly available corpus of transcribed Indian language speech developed for automatic speech recognition, and were used to supplement volunteer recordings for this affectively ambiguous class. The dataset comprises read news recordings from All India Radio, with audio provided in 16 kHz WAV format and aligned transcriptions in fairseq format.

The audiovisual component was assembled by extracting emotionally labelled utterances from Indian Punjabi films and television programmes,

providing naturally occurring speech with diverse speaker identities, recording conditions, and expressive styles not present in scripted volunteer data. While all audiovisual sources reflect Indian Punjabi, the volunteer component provides Pakistani Punjabi representation — all ten volunteer speakers were Pakistani Punjabi speakers. Although phonetic and prosodic differences between Indian and Pakistani Punjabi are not extensive, cross-border dialectal coverage remains uneven across the two components, and expanding audiovisual representation to Pakistani Punjabi content is an avenue for future work.

The final dataset comprises 893 labelled utterances used for all experiments reported in this paper.

4.2. Speaker Demographics

Table 1 summarises the demographic characteristics of the volunteer participant pool.

Table 1: Volunteer Speaker Demographics

Characteristic	Value
Total Speakers	10
Male Participants	5 (50%)
Female Participants	5 (50%)
Age Range	21–60 years
Native Punjabi Speakers	7 (70%)
Fluent Non-Native Speakers	3 (30%)
Utterance Duration	2–7 seconds

4.3. Class Distribution

The dataset covers four emotion categories: angry, happy, neutral, and sad. Table 2 presents the per-class sample counts broken down by source. Neutral samples were sourced exclusively from the Shrutilipi corpus (Bhogale et al., 2022) rather than audiovisual media, as neutral speech is rarely unambiguously expressed in dramatic contexts. The resulting class distribution is imbalanced, with sad utterances constituting the largest class (327) and happy the smallest (160). Such imbalance is an inherent property of naturalistic collection methodologies and introduces distributional priors that may interact with domain shift under cross-domain conditions.

4.4. RASA Dataset

The RASA dataset is a multilingual expressive Text-to-Speech (TTS) corpus (Varadhan et al., 2024) developed to support speech synthesis research across 22 Indian languages, including Punjabi. Each language subset features one male and one female speaker recorded under controlled, anechoic studio conditions. Audio is stored in mono WAV format at 48 kHz, with utterance durations ranging from 3 to 15 seconds. Despite its synthesis-oriented origins, the expressive nature of RASA recordings — spanning multiple emotion categories with consistent annotation — renders it suitable as a complementary SER resource, particularly as a controlled-domain counterpart to our naturalistic curated dataset.

Table 2: Emotional Class Distribution by Data Source

Emotion	Volunteer	AV Media	Shrutilipi*	Total
Angry	141	100	0	241
Happy	60	100	0	160
Neutral	0	0	165	165
Sad	233	94	0	327
Total	**434**	**294**	**165**	**893**

* Neutral samples sourced exclusively from the Shrutilipi corpus (Bhogale et al., 2022).

For this study, the Punjabi subset of RASA was used, covering the same four emotion classes as our curated dataset: angry, happy, sad, and neutral. RASA provides a predefined train–test split comprising 8,672 training samples and 962 test samples. The neutral class is heavily overrepresented, accounting for 6,039 of the 8,672 training samples. To mitigate this imbalance, two strategies were applied depending on the model architecture: for ResNet34, neutral training samples were capped at 2,000; for CNN-BiLSTM and Wav2Vec 2.0, class-weighted loss functions were used during training. The test partition was left unchanged in all cases to ensure fair evaluation.

Table 3 presents the class-level breakdown of the RASA Punjabi subset.

Table 3: RASA Punjabi Subset: Class Distribution

Emotion	Training Set	Test Set
Angry	861	95
Happy	915	102
Sad	857	95
Neutral	6,039	670
Total	**8,672**	**962**

The acoustic divergence between RASA's studio-controlled recordings and the naturalistic, variable-condition speech in our curated dataset provides the primary motivation for the cross-domain experimental framework. This contrast allows us to isolate the effect of domain mismatch from confounds such as language or emotion label differences, yielding a controlled setting for evaluating model generalisation.

5. Methodology

5.1. Pre-processing

All audio files were resampled to 16 kHz to ensure a uniform sampling rate across both corpora (Baevski et al., 2020). Each utterance was normalised to a fixed length of 16,000 samples (1 second at 16 kHz) via zero-padding for shorter clips and centre-cropping for longer ones. Processed files were structured into a Hugging Face `Dataset` object to standardise splitting, batching, and feature extraction across all experimental conditions.

5.2. Feature Extraction

Feature extraction followed a multi-representation strategy to capture complementary acoustic dimensions of emotional speech. The primary feature set comprised Mel-Frequency Cepstral Coefficients (MFCCs), which encode the short-term power spectrum of speech on a perceptually motivated scale (Gourisaria et al., 2024). Four additional feature families were extracted: chroma features, mel spectrograms, spectral contrast, and tonnetz features. Concatenating these descriptors into a unified feature vector provides complementary spectral, temporal, and harmonic information, improving discriminability across emotion classes that share prosodic surface characteristics.

5.3. Dataset Splitting

The splitting strategy varied by experimental strategy. For experiments using the curated dataset, an 80/20 stratified split was applied. The split was not enforced to be speaker-independent, and E1 results should be interpreted with this caveat in mind. For experiments involving RASA, no splitting was performed; the dataset was used in its entirety for either training or testing depending on the experimental strategy. Complete details of each configuration are presented in Table 4.

5.4. Model Architectures

Three architectures were evaluated across two model families. CNN-BiLSTM combines a convolutional front-end for local spectro-temporal feature extraction with a bidirectional LSTM for sequential modelling. ResNet34 applies 34-layer residual learning with skip connections; training used pitch shifting, time stretching, and temporal shifting augmentation, Adam optimisation with a discriminative learning rate (10^{-5} to 10^{-4}), and early stopping on validation loss. Wav2Vec 2.0 (`wav2vec2-xls-r-300m`) is a 300M-parameter transformer pretrained on large-scale multilingual speech (Baevski et al., 2020), fine-tuned here for four-class Punjabi emotion classification. All models were evaluated on accuracy, macro F1-score, precision, and recall.

5.5. Experimental Strategies

Four strategies were designed to evaluate in-domain performance, cross-domain transfer, and generalisation under domain shift — a dimension largely absent from prior Punjabi SER studies. Table 4 summarises each configuration, where "Curated" refers to our naturalistic Punjabi SER dataset.

Table 4: Experimental Training and Testing Strategies

Strategy	Training Set	Testing Set
E1	Curated	Curated
E2	RASA	Curated
E3	Curated	RASA
E4	RASA + Curated	Curated

E1 establishes in-domain baselines using the curated dataset for both training and testing. E2 and E3 probe cross-domain transfer in both directions between studio and naturalistic speech. E4 utilises a combined training set of both corpora to assess whether heterogeneous data can recover performance lost under domain shift.

6. Results

In this section, we evaluate the performance of the proposed models across the four experimental strategies (E1–E4) and analyze their effectiveness in recognizing emotions in Punjabi speech under both in-domain and cross-domain conditions.

6.1. CNN-BiLSTM

Table 5: CNN-BiLSTM Performance Across Experimental Strategies

Strategy	Macro Acc. (%)	Angry (%)	Happy (%)	Sad (%)	Neutral (%)
E1	85.39	100.0	80.0	66.7	93.94
E2	37.7	36.2	42.7	24.2	47.5
E3	34.0	54.6	0.0	29.4	51.8
E4	83.7	72.9	90.0	81.2	90.6

CNN-BiLSTM was evaluated for four-class emotion classification across all experimental strategies. Table 5 presents macro accuracy and per-emotion performance.

In the in-domain setting (E1), the model achieved high performance across emotions, with accuracies of 100.0% (angry), 80.0% (happy), 66.7% (sad), and 93.94% (neutral), resulting in a macro accuracy of 85.39%. This indicates that the model effectively

captures emotional patterns when training and testing distributions are aligned.

Cross-domain evaluation (E2) revealed a substantial drop in performance, with macro accuracy of 37.7%. While neutral achieved moderate recognition (47.5%), sad was poorly classified (24.2%), highlighting sensitivity to acoustic variability and domain mismatch between training and testing datasets.

Reverse transfer (E3) further exposed generalization limitations, yielding 34.0% macro accuracy. The model collapsed on happy (0.0%) and underperformed on sad (29.4%), demonstrating strong dataset-specific bias and insufficient domain-invariant learning.

Combined-domain training (E4) improved robustness considerably, achieving 83.7% macro accuracy. Performance became more balanced across emotions: angry (72.9%), happy (90.0%), sad (81.2%), and neutral (90.6%). Although slightly lower than the best in-domain result, this strategy substantially mitigated cross-domain degradation, demonstrating that heterogeneous training data enhances generalization in emotion recognition across diverse acoustic conditions.

6.2. ResNet34

Table 6: ResNet34 Performance Across Experimental Strategies

Strategy	Macro Acc. (%)	Happy (%)	Anger (%)	Sad (%)	Neutral (%)
E1	92.00	88.00	89.00	94.00	94.00
E2	37.36	27.33	62.92	40.80	1.85
E3	43.97	6.86	57.89	53.88	0.00
E4	96.00	100.00	89.58	100.00	93.93

ResNet-34 was evaluated across all four experimental strategies for four-class emotion classification. To mitigate overfitting on the curated Punjabi samples, we employed pitch shifting, time stretching, and time shifting augmentations. Optimization was conducted via the Adam optimizer using a discriminative learning rate slice (10^{-5} to 10^{-4}), with an Early Stopping mechanism to ensure convergence based on validation loss.

As shown in Table 6, the model achieved its peak performance under Strategy E4 (combined-domain training), reaching an overall accuracy of 96%. The model reached 100% accuracy for both the Happy and Sad categories, indicating that these emotions exhibit very clearly distinguishable acoustic patterns in the Punjabi corpus. A slight performance decrease was noted for Neutral (93.93%), while Anger proved challenging, at 89%. This relative drop in Anger likely reflects the acoustic overlap between high-intensity vocalizations and the noise present in the curated data.

The model exhibited its lowest performance in strategy E2, achieving an overall accuracy of 37%. In particular, it demonstrated substantial difficulty in correctly classifying neutral utterances. A comparable decrease in accuracy was observed in strategy E3; while slightly better at 43%, it failed to classify neutral entirely. Training details are provided in Table 7.

Table 7: ResNet34 Training Parameters

Parameter	Value
Model Type	ResNet34
Epochs	20
Batch Size	16
Optimizer	Adam
Learning Rate	Slice (10^{-5} to 10^{-4})
Loss Function	Flat Cross-Entropy

6.3. Wav2Vec 2.0

Table 8: Wav2Vec 2.0 Performance Across Experimental Strategies

Strategy	Macro Acc. (%)	Angry (%)	Happy (%)	Sad (%)	Neutral (%)
E1	92.48	100.0	87.0	86.0	100.0
E2	41.3	46.7	43.3	54.9	20.4
E3	25.0	0.0	0.0	100.0	0.0
E4	82.8	81.2	76.7	73.4	100.0

Wav2Vec 2.0 was fine-tuned for four-class emotion classification using contextualized speech representations derived from self-supervised pretraining. Table 8 reports macro accuracy along with per-emotion performance for all experimental strategies.

In the in-domain setting (E1), the model achieved strong and balanced performance across all emotions, with accuracies of 100.0% (angry), 87.0% (happy), 86.0% (sad), and 100.0% (neutral), resulting in a macro accuracy of 92.48%. This demonstrates that transformer-based representations effectively capture emotional characteristics when training and testing distributions are aligned.

Under distributional shift, performance degraded substantially. In E2, cross-domain transfer from RASA to curated speech yielded 41.3% macro accuracy, with moderate performance on sad (54.9%) but poor recognition of neutral (20.4%), highlighting the model's sensitivity to acoustic and recording variability.

Reverse transfer (E3) further exposed generalization limitations, with a macro accuracy of 25.0%. The model collapsed toward the sad class (100%) while failing to classify the remaining emotions, indicating strong dataset-specific bias rather than domain-invariant learning.

Combined-domain training (E4) improved robustness considerably, achieving 82.8% macro accuracy. Performance became more balanced, with

notable improvements in angry (81.2%) and happy (76.7%), and perfect recognition of neutral (100%). Although slightly lower than the best in-domain result, this strategy substantially outperformed pure cross-domain transfer, demonstrating that exposure to heterogeneous training data mitigates domain shift and enhances generalization.

7. Discussion

7.1. In-Domain Performance and the Value of Naturalistic Data

The strong in-domain results achieved across all three architectures in E1 — Wav2Vec 2.0 at 92.48% macro accuracy, ResNet34 at 92.00%, and CNN-BiLSTM at 85.39% — establish that naturalistic Punjabi speech data, even at modest scale, supports effective emotion recognition when training and test distributions are aligned. These results indicate that data ecological validity — the degree to which a dataset reflects real-world speech conditions — contributes substantially to in-domain model performance. The competitive performance of ResNet34 relative to Wav2Vec 2.0 further suggests that convolutional architectures with residual connections are capable of extracting stable spectro-temporal features from short naturalistic utterances, and that the representational advantage of transformer-based models is most pronounced under conditions of extreme data scarcity rather than when a well-distributed naturalistic corpus is available.

7.2. Cross-Domain Transfer and the Studio-to-Naturalistic Gap

The most significant finding of this study is the severity and consistency of performance degradation under cross-domain conditions. In both E2 (RASA $\rightarrow$ Curated) and E3 (Curated $\rightarrow$ RASA), all three models experienced dramatic reductions in macro accuracy. This degradation is attributable to two compounding factors. The primary source is the fundamental acoustic domain mismatch between RASA's controlled, anechoic studio recordings and the naturalistic recordings in our curated dataset, which encompass ambient noise, variable microphone conditions, and spontaneous emotional expression drawn from real-world audiovisual content. The second contributing factor is class imbalance, which amplifies cross-domain collapse by introducing strong distributional priors: the neutral class dominates RASA at 6,039 of 8,672 training samples, while *sad* constitutes the majority class in our curated set at 327 of 893 samples. Under domain shift, when cross-domain acoustic features become unreliable, models exploit these priors and converge toward the majority class of their training distribution — a well-documented failure mode in cross-corpus SER. The complete convergence of Wav2Vec 2.0 toward the *sad* class in E3 — achieving 100% sad recall while failing entirely on all other classes — exemplifies this mechanism precisely.

The directional asymmetry between E2 and E3 provides further insight: transfer from RASA to the curated set consistently outperforms the reverse direction, indicating that pretraining on larger, acoustically cleaner data confers a modest but measurable generalisation benefit. Nevertheless, this advantage remains insufficient to overcome the acoustic divergence between studio and naturalistic domains.

7.3. Combined-Domain Training as a Practical Strategy

E4 substantially recovers performance across all models, with ResNet34 achieving 96.00% and CNN-BiLSTM reaching 83.7%, indicating increased robustness to domain variation when heterogeneous training data is available. These results suggest that combining available controlled corpora with modest naturalistic data is more effective than relying on either source in isolation. The divergence between ResNet34 (96.00%) and Wav2Vec 2.0 (82.8%) in E4 is explained by the fine-tuning dynamics of each architecture. ResNet34 is trained from scratch on the combined corpus with task-specific augmentation — pitch shifting, time stretching, and temporal shifting — enabling full adaptation to the combined distribution. Wav2Vec 2.0 enters fine-tuning with deep representational priors from large-scale multilingual pretraining, which may not fully align with Punjabi emotional speech acoustics; when the fine-tuning corpus remains small relative to pretraining scale, these priors can hinder full adaptation to the target domain.

7.4. Broader Implications for South Asian SER

These findings collectively establish that acoustic domain match is a key determinant of model performance in Punjabi SER, in this setting often exerting greater influence than architectural sophistication or pretraining scale. Evaluation benchmarks that report exclusively in-domain performance may overstate model generalisability — a concern of particular consequence when studio corpora constitute the primary available resource for South Asian languages. The cross-domain evaluation framework introduced in this study provides a more diagnostically informative paradigm that can complement standard in-domain assessment in future work. Furthermore, the E4 results demonstrate that small naturalistic corpora provide substantial gains

when combined with larger controlled datasets — a tractable data collection strategy for South Asian language communities operating under resource constraints. Given the phonetic and prosodic proximity of Punjabi to Hindi and Urdu, the cross-domain evaluation framework introduced here may generalise naturally to these related languages, and future work exploring multilingual transfer across South Asian languages — including whether Hindi or Urdu training data can improve Punjabi SER performance — would position this line of research within a broader coordinated regional effort.

8. Error Analysis

8.1. Emotion-Level Confusion Patterns

Examination of per-class performance reveals consistent patterns across models and strategies that macro accuracy alone does not capture. In the in-domain setting (E1), angry achieved perfect precision and recall (100%) across both CNN-BiLSTM and Wav2Vec 2.0, reflecting its acoustically distinctive profile — elevated pitch, high energy, and rapid articulation. ResNet34 similarly performed strongly in E1, with the only notable errors being 3 angry utterances misclassified as sad and 3 happy utterances misclassified as angry, consistent with acoustic overlap between high-intensity emotions. The clearest in-domain weakness was sad: CNN-BiLSTM recorded only 66.67% recall despite 100% precision, misclassifying 5 sad utterances as neutral and 3 as happy. This confusion is consistent with the acoustic proximity between subdued sadness and neutral speech, both characterised by reduced pitch range, lower energy, and slower speaking rates. Wav2Vec 2.0 showed a complementary pattern, with happy (87% recall, 77% precision) as its primary weakness — predominantly misclassified as neutral — suggesting that low-arousal happy expressions in naturalistic Punjabi share prosodic characteristics with neutral utterances.

Under cross-domain conditions, these confusions intensified severely. In E2 (RASA → Curated), CNN-BiLSTM's sad recall collapsed to 24.2%, while Wav2Vec 2.0 showed the inverse pattern: highest sad recall (54.9%) but near-collapse on neutral (20.4%). ResNet34 under E2 exhibited a striking neutral collapse — of 162 neutral test instances, 156 were misclassified as sad, mirroring the majority-class prior effect observed in Wav2Vec 2.0 under E3. In E3 (Curated → RASA), Wav2Vec 2.0 assigned sad to all test instances — achieving 100% sad recall and 0% on all other classes — a complete distributional collapse driven by the majority-class prior of the curated training set interacting with the acoustic divergence of studio test conditions. CNN-BiLSTM in E3 exhibited a comparatively distributed failure, with happy collapsing entirely (0%) while angry (54.6%) and neutral (51.8%) retained partial recognition. ResNet34 under E3 showed a distinct failure mode: sad recalled 0%, with all sad instances redistributed to angry (20) and neutral (75), while angry and neutral retained partial recognition.

Combined-domain training (E4) substantially restored class balance across all three models. CNN-BiLSTM exceeded 72% recall on all four classes, with happy (90.0%) and neutral (90.6%) recovering most strongly. Wav2Vec 2.0's sad (73.4%) remained its relative weak point while neutral achieved perfect recognition (100%) — a complementary profile to CNN-BiLSTM that points toward ensemble approaches as a promising direction for further gains. ResNet34 in E4 achieved the most balanced recovery of all models, with near-perfect classification across all four classes and minimal inter-class confusion, consistent with its 96% macro accuracy.

8.2. Cross-Model Error Consistency

A consistent finding across all three models and all strategies is that angry is the most robustly recognised emotion and that neutral recognition is the most fragile under domain shift. The happy–sad pair remains the most frequently confused under in-domain conditions, with both emotions capable of manifesting with moderate pitch and tempo in naturalistic Punjabi depending on speaker register — a fundamental acoustic ambiguity that warrants targeted dataset design in future work. Table 9 summarises the dominant misclassification patterns across all models and strategies.

Table 9: Dominant Misclassification Patterns Across Models and Strategies

Strategy	Model	Primary Confusion	Secondary Confusion
E1	CNN-BiLSTM	Sad → Neutral	Happy → Neutral
E1	ResNet34	Angry → Sad	Happy → Angry
E1	Wav2Vec 2.0	Happy → Neutral	Sad → Happy
E2	CNN-BiLSTM	Sad → Neutral	Angry → Neutral
E2	ResNet34	Neutral → Sad	Happy → Angry
E2	Wav2Vec 2.0	Neutral → Sad	Angry → Sad
E3	CNN-BiLSTM	Happy → Neutral	Sad → Neutral
E3	ResNet34	Sad → Neutral	Happy → Angry
E3	Wav2Vec 2.0	All → Sad	—
E4	CNN-BiLSTM	Angry → Sad	Sad → Neutral
E4	ResNet34	Angry → Sad	—
E4	Wav2Vec 2.0	Sad → Happy	Happy → Sad

9. Comparative Analysis

Table 10 situates our results within the existing Punjabi SER literature. Direct numerical comparison across studies is complicated by differences in dataset composition, emotion label sets, and evaluation methodology; however, the table estab-

lishes approximate performance trajectories across successive approaches.

Table 10: Comparison with Prior Punjabi SER Work

Study	Dataset	Best Model	Accuracy (%)
Singla & Singh (2022)	985 utterances (film/TV)	Random Forest	61.65
Singla et al. (2024)	9,000 utterances (film/web)	CNN	69.00
Kaur & Singh (2022)	900 utterances (recorded)	CNN	81.00
This work (E1)	**893 naturalistic**	**Wav2Vec 2.0**	**92.48**
This work (E4)	**RASA + 893 naturalistic**	**ResNet34**	**96.00**

Our in-domain result of 92.48% (Wav2Vec 2.0, E1) represents a substantial improvement over the prior state-of-the-art for Punjabi SER. Singla & Singh (2022) achieved 61.65% using classical feature engineering with Random Forest on semi-natural film speech, while Singla et al. (2024) advanced this to 69% using CNNs on a considerably larger 9,000-utterance corpus. Kaur & Singh (2022) further advanced this to 81.00% using a 1D CNN with LASSO feature selection on a purpose-recorded 900-utterance corpus, demonstrating the benefit of dedicated data collection over media-sourced datasets. Our results demonstrate that transformer-based self-supervised representations close this gap significantly, achieving over 23 percentage points of improvement relative to Singla et al. (2024) on a smaller but more acoustically diverse dataset. The combined-domain result of 96.00% (ResNet34, E4) further extends this margin, suggesting that data heterogeneity is as consequential as model architecture in advancing Punjabi SER performance.

Critically, prior Punjabi SER studies evaluate exclusively in-domain, leaving cross-domain generalisation entirely uncharacterised. This study introduces the first systematic cross-domain evaluation framework for Punjabi SER, establishing baseline cross-domain performance figures — E2: 37.7–41.3%, E3: 25.0–34.0% — that reveal the brittleness of current approaches under domain shift and provide a reproducible benchmark against which future work can be measured.

10. Conclusion

This paper presented a systematic cross-domain evaluation of deep learning and transformer-based models for Punjabi Speech Emotion Recognition. We introduced a curated dataset of 893 naturalistic utterances and evaluated CNN-BiLSTM, ResNet34, and Wav2Vec 2.0 across four experimental strategies — in-domain (E1), cross-domain transfer (E2), reverse transfer (E3), and combined-domain training (E4) — characterising model performance under both aligned and mismatched distribution conditions.

Three principal findings emerge. First, Wav2Vec 2.0 achieves 92.48% macro accuracy in-domain, substantially advancing the prior Punjabi SER state-of-the-art. Second, cross-domain transfer degrades severely across all models and both transfer directions, demonstrating that in-domain evaluation alone is insufficient to assess practical model utility. Third, combined-domain training recovers performance effectively — ResNet34 reaching 96.00% — establishing data heterogeneity as the most practical development strategy for low-resource Punjabi SER. This work further contributes the first cross-domain evaluation framework for Punjabi SER, providing reproducible baselines and motivating evaluation protocols that assess domain generalisation alongside in-domain accuracy. Future work will prioritise dialectal dataset expansion to cover Pakistani Punjabi audiovisual content, domain adaptation techniques for studio-to-naturalistic transfer, multilingual transfer learning across South Asian languages including Hindi and Urdu, and extension to dimensional emotion recognition.

Ethics and Limitations Statement

All volunteer participants provided informed consent prior to recording. Participants were informed of the research purpose and intended use of their recordings, and participation was voluntary. Audiovisual media samples were used solely for non-commercial academic research. No personally identifiable information beyond basic demographic characteristics is retained. Speech emotion recognition systems carry inherent risks of misuse in surveillance and non-consensual affect monitoring; the authors do not endorse such applications, and the resources developed in this work are intended strictly for research use under appropriate ethical oversight.

This work has several limitations. First, the curated dataset comprises 893 utterances from a small volunteer pool of ten speakers, and the 80/20 train-test split was not enforced to be speaker-independent; E1 results should therefore be interpreted with caution as they may reflect some degree of speaker overlap. Second, all audiovisual media sources reflect Indian Punjabi, while volunteer speakers represent Pakistani Punjabi; cross-border dialectal coverage therefore remains uneven, and generalisation to Pakistani Punjabi audiovisual content is untested. Third, the dataset does not represent the full dialectal diversity of Punjabi as spoken across India and Pakistan, and findings may not generalise uniformly across all regional varieties. Finally, the cross-domain results establish that models trained on studio-controlled corpora generalise poorly to naturalistic speech, suggesting that the high in-domain accuracies reported here and in prior work overstate practical deployability.

Acknowledgements

The authors sincerely thank all volunteer participants for contributing their recordings to this study. We also acknowledge the providers of publicly available datasets used in this work, whose resources made this research possible. Their contributions are greatly appreciated.

Data and Code Availability Statement

The dataset is publicly available on Kaggle. The code is available on this repository.

11. Bibliographical References

[1] Aggarwal, S., Singh, P. & Sharma, K. (2023). A survey on speech emotion recognition: Datasets, features, and classification methods. *Journal of Intelligent Systems*, 33(1), pp. 203–220.

[2] Baevski, A., Zhou, H., Mohamed, A. & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*.

[3] Barhoumi, C. & Ben Ayed, Y. (2023). Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation. *Research Square* [Preprint]. doi: 10.21203/rs.3.rs-2874039/v1.

[4] Bhogale, K. S., Jain, N., Raman, A., Agarwal, S., Chhimwal, A., Bansal, D. & Jyothi, P. (2022). Effectiveness of Mining Audio and Text Pairs from Public Data for Improving ASR Systems for Low-Resource Languages. *arXiv preprint arXiv:2208.12666*.

[5] El Ayadi, M., Kamel, M. S. & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), pp. 572–587.

[6] Gourisaria, M. K., Agrawal, R., Sahni, M. & Singh, P. K. (2024). Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. *Discover Internet of Things*, 4(1).

[7] Kaur, K. & Singh, P. (2022). Impact of Feature Extraction and Feature Selection Algorithms on Punjabi Speech Emotion Recognition Using Convolutional Neural Network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(5), Article 100.

[8] Li, Y., Wang, Y., Yang, X. & Im, S. (2023). Speech emotion recognition based on Graph-LSTM neural network. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(40).

[9] SIL International. (2023). Punjabi Language Profile. *Ethnologue*. [Online]. Available: https://www.ethnologue.com/language/pan

[10] Singla, C. & Singh, S. (2022). Punjabi Speech Emotion Recognition using Prosodic, Spectral and Wavelet Features. In *Proceedings of the 2022 10th IEEE International Conference on Emerging Trends in Engineering & Technology Signal and Information Processing (ICETET-SIP-22)*, pp. 1–6.

[11] Singla, C., Singh, S., Sharma, P., Mittal, N. & Gared, F. (2024). Emotion Recognition for Human–Computer Interaction Using High-Level Descriptors. *Scientific Reports*, 14, p. 12122.

[12] Varadhan, P. S., Sankar, A., Raju, G. & Khapra, M. M. (2024). Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Low-resource Settings. In *Proceedings of INTERSPEECH 2024*.

[13] Ververidis, D. & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features and methods. *Speech Communication*, 48(9), pp. 1162–1181.

12. Language Resource References

[1] Singla, C. & Singh, S. (2022). PEMO: A New Validated Dataset for Punjabi Speech Emotion Detection. *International Journal on Recent and Innovation Trends in Computing and Communication*, 10(10), pp. 52–57. ISLRN 2321-8169.