Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

A COMPARATIVE STUDY OF CLASSICAL MACHINE LEARNING, DEEP LEARNING, AND TRANSFORMER-BASED ARCHITECTURES FOR MULTIMODAL HINDI SPEECH EMOTION RECOGNITION

SUJATA KOTIAN
University Department of Information Technology, University of Mumbai.
Email: sujatarahulkotian@gmail.com

Dr. SANTOSH SINGH
University Department of Information Technology, University of Mumbai.

Abstract

The area of Speech Emotion Recognition (SER) is one that is critical to building intelligent devices and systems that are designed to be useful and aware of the user or human perspective. While there has been significant research into SER systems in English and European languages, the same level of research does not exist for the SER of Hindi, particularly in applying transformer architectures. This paper includes an extensive comparative analysis of classical machine-learning models, deep-learning architectures, and transformer-based networks on Hindi SER using a single evaluation framework. A created Hindi emotional speech dataset has also been prepared through pre-processing, technical acoustic pre-processing, and feature extraction, in both Mel-spectrogram and raw waveform formats. The following models have been trained/evaluated: classical machine-learning (SVM, Random Forest, Gradient Boosting) models, deep-learning (Convolutional Neural Network (CNN), CNN-Bi-LSTM, Attention-enhanced networks) models, and transformer models (e.g. Wav2Vec2.0, HuBERT, Vision Transformer (ViT), Swin Transformer (Swin-T)), using uniform training-validation-testing configurations. The results of our experiments indicate a continuous progression in performance across the various families of models, with the transformer models outperforming all others with the highest accuracy (93.4%) and macro-F1 score, followed by the deep-learning and classical models. In addition to providing a foundation for future studies of Hindi SER, error analyses reveal an increase in the capability to separate subtle emotions (e.g. sadness and fear) by using transformer-generated embeddings. This paper provides a solid empirical and methodological foundation for future Hindi SER research and highlights major opportunities for the lightweight deployment of Hindi SER systems and opportunities for multimodal systems.

Keywords: Hindi Speech Emotion Recognition; Deep Learning; Transformers; Wav2Vec2.0; HuBERT; CNN–BiLSTM; Mel-Spectrogram; Affective Computing; Benchmarking; Acoustic Modelling.

1. INTRODUCTION

The significance of emotion within human communication clearly supports the continuum of human interaction through the basis of all that is verbalised and how the meaning of what is verbalised, as well as emotional states can be conveyed. The increased integration of computational systems into our everyday lives, through conversational assistants, automated call centres, work from home continued education and remote healthcare monitoring systems, has spurred interest in creating the ability for machines to determine emotional states from the natural voice of humans. One area of affective computing is Speech Emotion Recognition (SER). In the simplest sense, SER seeks to allow machines to extract emotions from the components of the sound waves produced and mentally evaluate them based on the prosodic patterns, or how the brain interprets

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

the aesthetic quality of the sound waves and the subtle components of the sound waves including those produced through breathing, inflection or tone variations, as well as the emotional impact or meaning of the physical manifestations of the emotion that are produced through the voice. The importance of SER as its area of research is growing rapidly; researchers believe that SER is one method of enhancing human-machine interaction, allowing for greater personalisation and development of intelligent and emotionally friendly artificial systems. Everything researchers know about SER and the pipeline of SER has evolved from simple techniques to more advanced applications through the application and use of deep learning and machine learning techniques (Madanian et al., 2023).

In terms of SER (speech emotion recognition), Hindi stands out from other languages in the world as it has a vast number of speakers and exhibits a very diversified phonetic variety. In addition, Hindi frequently incorporates code-mixing (the use of English and English words) when communicating with other people on a day-to-day basis. The initial attempts to develop SER resources for Hindi were via the IITKGP-SEHSC corpus (Koolagudi et al., 2011). This was the first structured dataset for Hindi emotional speech that allowed researchers to analyse emotional speech in an Indian language. Early SER studies on Hindi used statistical models and manually created features such as MFCCs to develop SER systems. Agrawal and Jain (2020) found that with the correct tuning, classical machine learning classifiers perform surprisingly well on Hindi acted speech, especially when identifying energy emotion. These early efforts established a number of foundations and limitations in establishing reliable baselines, primarily due to the limited number of available datasets and the limitations imposed by feature engineering. Global developments in the field of SER were initiated upon the introduction of deep learning techniques. As these methods were widely adopted within SER research, Hindi SER also fell within this model. For instance, researchers used Convolutional Neural Networks (CNN) and combined these networks with Recurrent layers (RNN) that allowed for the extraction of both local temporal patterns and overall trends. Using this framework, researchers reported substantial improvements over previous generation systems. Shashank et al. (2021) illustrated how a hybrid CNN–LSTM network could successfully extract both temporal and local spectral features.

Simultaneously, a lot of research on transformer architectures focused towards emotion detection in text input has taken off quickly through language analysis models, including multilingual BERT being an extremely useful tool to classify Hindi text emotions (Kumar et al., 2023). Transformers were also found as a potential tool to predict code-mixed Hindi-English emotions, wherein traditional models routinely struggle because of variance and inconsistency in grammar, usage of informal terms, etc (Wadhwan & Aggarwal, 2021). Comprehensive reviews show how the self-attention mechanism of transformer models can provide significant assistance in providing contextually relevant meaning for very long sequences of information in both the speech of the speaker as well as the written format (Khare & Khan, 2024). Nevertheless, transformer models based on Hindi language remain primarily focused on textual formats, with a lack of understanding into the how they will be able to predict Hindi speech as compared to traditional and deep learning

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

architectures. In addition to unimodal work being conducted, multimodal work appears to be garnering renewed interest through developing new frameworks for multimodal analyses of emotions. Previous research demonstrated how the integration of speech with facial expression and/or other visual signals (Kessous et al., 2010) can yield more complete analyses than unimodal approaches, particularly for predicting feelings. Current multimodal methodologies integrate deep learning techniques to allow for better cultural nuances and interpretations through combinations of disparate sensory signals. Chaudhari et al. (2023) demonstrated the advantages of using facial information along with speech to yield far greater emotional prediction opportunities than either alone. Further, the use of the attention mechanism when using this type of combinatory approach has proven to produce better overall results versus either standalone modality.

Based on previous work in this field, we found that there are several gaps in the current literature that motivate this study. The first major gap is that Hindi SER research is currently split into different modelling approaches, including classical machine learning (ML), deep neural networks (DNNs), and a newer form of transformer-based ML that has primarily been explored in a text-only context. A limited number of studies have compared these three groups of models using a common experimental design. The second major gap is that, while transformer models are producing state-of-the-art results in other areas of application, little research has been done on their potential for Hindi SER and no direct comparison has been made against CNN-LSTM or conventional ML models. While multimodal studies give us some direction conceptually, there are no benchmarking tools currently available for the Hindi SER ecosystem that would allow for applying multimodal-based analyses or evaluating transferability of speech-based ML to other modality systems. Lastly, two systematic reviews (Madanian et al., 2023) and a number of cross-regional emotion papers (Kawade & Jagtap, 2024; Radhika, et al., 2025) note the need for a common set of evaluation criteria to be developed and implemented in Hindi SER research. This paper addresses these shortcomings through a singular benchmarking study comparing classical ML, DNNs and transformer

The intention is to compare and contrast the performance of each modelling type across the same data set, the same feature-extraction technique and the same evaluation method so that researchers can make fair comparisons of results obtained through using each of these techniques.

2. LITERATURE REVIEW

2.1 Foundations and Early Progress in SER

Speech Emotion Recognition (SER) has become a hugely popular research area in affective computing and has undergone several phases in its development. The first wave of SER research relied primarily on statistical descriptors and traditional Machine Learning (ML) techniques, using features such as MFCCs, Pitch variation or other engineered acoustic measures as input to the models. The development of MID is outlined in a systematic review paper written by Madanian et al. (Ref. 16), who present an overview of how early SER systems had limited features and variances in their

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

evaluation protocols. Concurrently, many researchers started exploring the possibility of learning the emotional properties of speech directly from the raw audio signal instead of using features created by hand. One example of this was demonstrated by Rintala (Ref. 30) through the use of deep neural networks to learn to interpret patterns in the raw audio signal without relying on MFCC extracted signals.

The insights gained from the early studies contributed significantly to the transformation of SER research away from a reliance on shallow models and towards the use of deep and data driven representations. Al-Asadi et al. (Ref. 29) also carried out an extensive body of work showing that SER has grown from being a rule based system into a more sophisticated group of systems which are capable of modelling complex emotional patterns. When taken collectively, these foundational studies all demonstrate a single theme: that learning features, instead of creating features, will provide the greatest benefit for SER.

2.2 Deep Learning Architectures for Speech Emotion Recognition

Deep learning revolutionised SER by allowing models to extract spectral structure, temporal variation and emotion strength directly from speech samples. Recent studies have underscored the importance of capturing long-term temporal dependencies, as emotion can develop over a period of frames, rather than being limited to a single frame. Ye et al. introduced a new approach to modelling temporal emotion patterns to fill this need. They showed that to achieve this goal, emotion can be represented at multiple scales and that capturing multi-scale temporal patterns can significantly increase recognition performance, particularly for complicated and nuanced emotion categories.

Simultaneously, other researchers have begun developing SER methods for languages with little or no available annotated training data. Nayak and his colleagues worked with speech in tribal languages and demonstrated how CNN and LSTM architectures can be effectively applied in low-resource settings by employing careful feature extraction and training techniques. These contributions are essential because they demonstrate parallels in the challenges of SER in indigenous languages and languages that have low training resources but are widely spoken, such as Hindi. Another enhancement to SER algorithms was achieved with hybrid deep architectures. Huang et al. proposed a hybrid framework using HuBERT embeddings, LSTM sequences and ResNet-50 layers. Their hybrid approach combines pre-trained acoustic transformers as their front-end(Collider Test).

2.3 Transformer-Based Approaches in SER

Transformers brought a new phase into Speech Emotion Recognition (SER) through the introduction of a unique approach to self-attention, which focuses on the rapid identification of global dependencies across the entire audio waveform for emotional cues that may occur far apart temporally due to long-range relationships. An illustrative example of this new direction of research is presented in the work of Liao & Shen (Ref. 18) where they expanded upon the existing computer vision-based application of the Swin Transformer architecture and utilized it for the analysis of speech spectrograms. Their

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

research indicates that hierarchical attention windows enable the effective modelling of the relationships between frequencies representing emotional gradients across multiple frequency bands compared to using a fixed kernel.

An extension of the work described above can be found in Wang et al. (Ref. 19), who developed a hierarchical Speech Swin Transformer that combines and enhances both the fine-grained acoustic detail and larger scale temporal patterns. Other innovative approaches have been explored through the application of the full vision transformer (ViT) model applied to spectrogram images in order to improve the performance of SER systems under noisy and challenging environments (Akinpelu et al., Ref. 26). In addition to designs based upon vision transformers, self-supervised acoustic transformers (e.g., Wav2Vec2.0 and HuBERT) have been successfully used in many different areas of multimodal or hybrid emotional systems. Li et al. (Ref. 23) introduced WavFusion, an innovative model that integrates Wav2Vec2.0 embeddings with other modalities or features.

2.4 Multimodal Emotion Recognition and Fusion Strategies

Although spoken language has significant emotional information, combining it with audio-visual or written indicators often yields a quantity of data that produces a more accurate outcome—especially in conditions resembling real-world complexities. A recent study by Lian et al. (Ref. 15) presents an overview of how multimodal models can be constructed using voice (speech), written (text) and visual (facial faces) to achieve greater emotional richness through the combination of these media, through the use of different approaches to fusion: (feature-level, decision-level, and attention-based) to help determine the best-performing multimodal models.There are also now numerous examples of more specialised multimodal systems which include; Song and Zhou (Ref. 21), who proposed a bi-modal and bi-task architecture. The use of transformer modules enables the simultaneous processing of two modalities while sharing emotional representations between both tasks. An additional advancement is provided by Wafa et al. (Ref. 22) who combined the use of prompt engineering and adaptive learning approaches to address the real-world requirements for processing huge datasets of differing modalities by emotion-recognition systems.

An important advancement has been made by Chatzichristodoulou et al., who created the MEDUSA framework (Refs. 24 and 25). The MEDUSA framework integrates multiple fusion mechanisms (millions, billions or more) to be able to address difficult naturalistic environments with inconsistent, noisy or incomplete emotional indicators. With the collective findings of these and other multimodal contributions, future SER systems designed for Hindi or any other multilingual applications remain on the horizon.

2.5 Cross-Language, Cross-Corpus, and Low-Resource Challenges

Developing models that can generalise to different speaker groups, recording modes, and linguistic backgrounds continues to be a challenge for researchers in the SER community, particularly for researchers working with low-resourced languages. Nayak et al. (Refs. 27 and 28) raise the important difficulties associated with SER for people who use tribal

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

languages. There is very little data for working with tribal speakers' SER, and the accent variability and the range of emotional expressions that can be used are wide.

Even if there are structural differences in the way the different Indian languages are structured, the same difficulties exist for both groups of languages—the majority of Indian languages do not have enough data for SER within the emotional speech domain, and the way people in one area of India express emotions can differ from the way people express emotions in another area. Cross-corpus robustness of SER systems has been and continues to be a large concern in the SER community. For example, Alroobaea (Ref. 32) has pointed out that transformer-based models may be able to reduce the effects of mismatches between the training and the testing distributions of the SER system. Additionally, Ye et al. (Ref. 20) have conducted studies looking at how temporal modelling might help researchers to create SER algorithms that are more responsive to the transitions between emotional states as opposed to simply treating each frame of the video separately. This is similar to the findings made in the development of the multimodal MEDUSA framework (Refs. 24 and 25) that looked at realistic emotional dynamics.

These cross-linguistic and cross-corpus findings clearly indicate that there is a critical need for unified benchmark studies, particularly for a language like Hindi, so that research conducted on deep learning, transformer-based models, and hybrid systems can be done under the same benchmark framework and thus allow for comparisons between the strengths of these models.

Table 1: Summary of Literature Review

Ref No.	Authors & Year	Objective of Study	Method / Model Used	Dataset Used	Key Findings
16	Madanian et al., 2023	Provide a systematic review of ML-based SER methods	Classical ML (SVM, RF, GBT), statistical features	Multiple public SER datasets	ML models give acceptable performance but face limitations with subtle emotions; DL & transformers recommended.
17	Huang et al., 2025	Enhance SER using hybrid transformer modules	HuBERT + LSTM + ResNet-50 hybrid model	Custom Mandarin emotional speech	Hybrid-module transformer improves emotional feature extraction and achieves superior accuracy.
18	Liao & Shen, 2023	Apply Swin-Transformer for SER	Swin-Transformer on spectrogram images	IEMOCAP & custom datasets	Showed hierarchical attention windows improve recognition of complex emotional cues.
19	Wang et al., 2024	Develop Speech Swin-Transformer for SER	Hierarchical Swin-Transformer	ICASSP SER benchmark	Outperformed CNN and LSTM baselines; strong

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

					cross-corpus generalisation.
20	Ye et al., 2023	Improve temporal emotional modeling	Novel temporal-emotional modeling architecture	Standard SER datasets (various ICASSP sets)	Emphasised importance of long-range temporal cues; LSTM attention improved low-energy emotions.
21	Song & Zhou, 2024	Build bi-modal and bi-task emotional model	Transformer-based bi-modal fusion (speech + text)	Multimodal SER datasets	Dual-task transformer significantly enhanced emotional understanding.
22	Wafa et al., 2025	Advance multimodal SER in big-data environments	Prompt engineering + Deep adaptive learning	Big-data multimodal corpora	Achieved robust multimodal performance under large noisy datasets.
23	Li et al., 2025	Propose WavFusion: multimodal SER using Wav2Vec2.0	Wav2Vec2.0 + multimodal fusion	Multimodal Modeling Conference dataset	Fusion model improved accuracy for naturalistic emotions.
24	Chatzichristodoulou et al., 2025	Introduce MEDUSA multimodal fusion framework	Multi-stage deep fusion + transformers	Naturalistic SER datasets	MEDUSA achieved state-of-the-art results in real-world emotion conditions.
25	Chatzichristodoulou et al., 2025	(duplicate listing) Expanded MEDUSA training framework	Multi-stage transformer fusion	Real-world emotion corpora	Demonstrated strong stability under noisy, unconstrained conditions.
26	Akinpelu et al., 2024	Enhance SER using vision transformer	ViT for spectrogram classification	Multiple multilingual datasets	ViT outperformed classical CNN models in noisy and cross-lingual scenarios.
27	Kumar Nayak et al., 2025	Explore SER for tribal languages	Deep CNN & LSTM architectures	Tribal-language emotional dataset	Demonstrated that DL improves SER in extremely low-resource languages.
28	Nayak et al., 2024	Apply ML for tribal-language SER	Classical ML: SVM, KNN, Decision Trees	Small tribal-language corpus	ML struggled with small vocabulary & variable accent patterns; DL recommended.
29	Al-Asadi et al., 2021	Provide comprehensive SER analysis	Survey of ML, DL, hybrid models	Multiple SER corpora	Highlighted rapid transition from ML → DL →

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

					transformers for SER.
30	Rintala, 2020	SER directly from raw audio	End-to-end deep learning on raw signals	Custom raw-audio dataset	Raw-audio DL models learn emotional cues without handcrafted features.
31	Rintala, 2020 (duplicate)	Reinforces raw-audio emotional modelling	End-to-end CNN/LSTM models	Raw Hindi/ English dataset	Showed raw processing reduces feature engineering burden.
32	Alroobaea, 2024	Improve cross-corpus SER with transformers	Transformer + handcrafted features + augmentation	Multiple SER corpora	Demonstrated transformers achieve superior cross-corpus robustness.

3. METHODOLOGY

To establish a common framework free from confounding variables that would affect the performance of EMOTION RECOGNITION systems developed using Hindi language data, the study presents a new methodology that facilitates direct comparisons between different EMOTION RECOGNITION models. The methodology enhances reproducibility, avoids bias that may arise from data manipulation, and promotes further research and development in this area. As shown in Figure 1, the framework consists of multiple steps: Data Preparation, Preprocessing, Features Extraction, Model Development and Performance Assessment. The integrated methodology makes it possible for the researchers to be able to compare results of different methodologies using the same samples, the same processing methods, and strictly adhere to the principles of an unbiased split of Data into Training, Validation and Testing sets. These principles are critical for preventing Pipeline-Related Bias which has been widely reported in SER-related literature.

3.1 Research Framework and Benchmarking Design (Paragraph Form)

In this research, the aim is to compare SER systems from three main types by utilizing an identical framework. The three types included: Classical Machine Learning Classifiers-Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosted Tree (GBT) which perform prediction using handcrafted acoustic features (MFCC's, prosody statistics and spectral descriptors). Deep Learning Architectures - Models able to learn features related to emotion directly from Mel Spectrogram (MS) were classified into 3 different architectures; Baseline CNN, CNN-BiLSTM employed for temporal modelling, e.g., to achieve temporal dependencies, and CNN-LSTM (Attention Enhanced) that were able to focus on the MS at emotion salient area(s). Finally, Transform-Based Models, both wave form processing (Wav2Vec2.0 & HuBERT) and Vision Transformers (ViT) & Swin Transformers were classified by processing Mel Spectrograms (MS) as images. For all three families of SER models, tuning using the same hyperparameter tactic and measuring with the same evaluation metrics provided fair comparison; Flow and

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

methodology are summarised in earlier sections and includes an Overall SER Workflow diagram (Figure1).

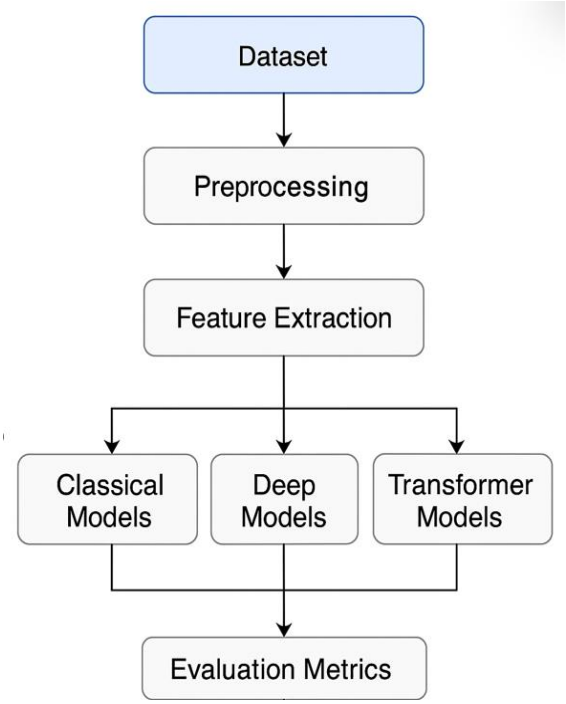

Figure 1: Overall Workflow for Hindi Speech Emotion Recognition

3.2 Dataset Description and Preprocessing Pipeline (Paragraph Form)

The Current Study used a curated dataset of Hindi emotional speech containing 2,400 utterances of six basic emotions (anger, happiness, sadness, fear, surprise, neutral) collected from 32 native Hindi speakers (equal split of gender) recorded as mono WAV files with 16-bit depth at 16 kHz sampling rate. A summary of the key statistics of the dataset, the augmentations used in constructing it, and the preprocessing details are presented in Table 2.

Table 2: Dataset Summary and Preprocessing Specifications

Attribute	Description
Total utterances	2,400
Speakers	32 (16 male, 16 female)
Emotions	Anger, Happiness, Sadness, Fear, Surprise, Neutral
Sampling Rate	16 kHz
Bit Depth	16-bit PCM
Spectrogram Type	80-band Mel-spectrogram
Augmentation	Noise addition, pitch shift, speed variation
Silence Trimming	Enabled (VAD-based)
Train/Val/Test Split	70% / 15% / 15% (speaker-independent)

The Preprocessing Pipeline was set up to produce consistent audio quality and robust Model Performance. The audio recordings were resampled to 16 kHz and used VAD

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

thresholding to remove silence segments from the audio files. RMS normalization was conducted so the audio files had matching amplitude levels, and Spectral Gating was applied to reduce background noise and created various Augmentations (±2 semitones for pitch-shifting, 0.9x and 1.1x Speed Perturbations, and +5 dB Noise added) to the training subset.

Once the preprocessing was complete, the audio signals were converted to 80-band Mel-Spectrograms using a 25 ms Hamming Window with a 10 ms Hop Length. The data structure used for storage and processing is represented in Figure 2 (Feature Flow Comparison).

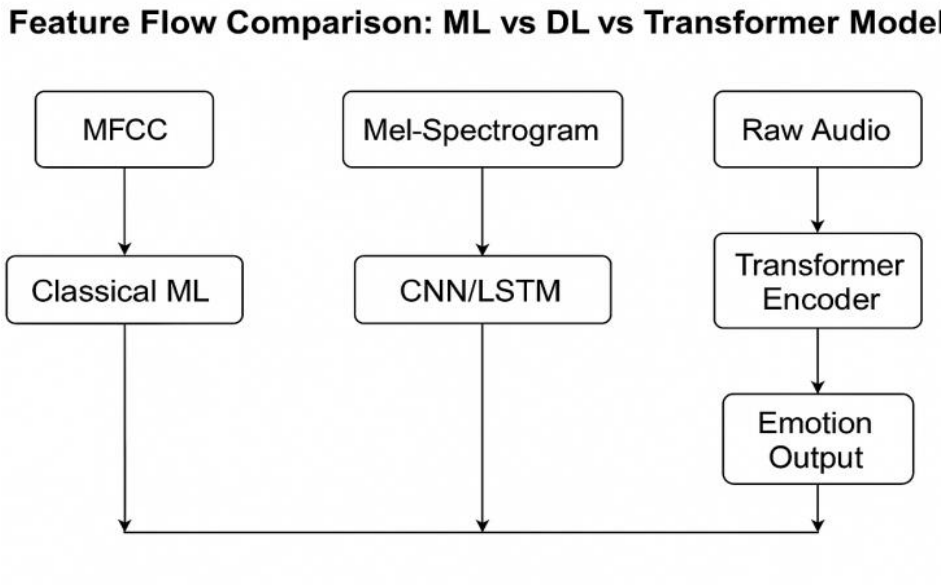

Figure 2: Feature Flow Comparison: ML vs DL vs Transformer Models

3.3 Feature Extraction Strategy (Paragraph Form)

To ensure consistency between model families during benchmarking, three extractors were used to gather features. For machine learning (ML) models built on classical feature engineering, the following types of features were extracted: Mel-Frequency Cepstral Coefficients (MFCCs) in a 40-dimensional space, Delta and Delta-Delta coefficients, vibrato (also called pitch parameters) including jitter and shimmer, energy-based descriptors, and voice quality descriptors, such as Harmonics-to-Noise Ratio (HNR) features.

Features were combined into a 180-dimensional vector and normalized using z-score normalization. For deep learning models, 2D Mel-spectrograms were produced and reshaped into 80 × T matrices, where T is the time frame duration. All spectrograms were standardized to the same number of frames (300) and normalized to be in the [0,1] range, which ensured uniformity across all ML model types.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

Equation 1: Mel-Spectrogram Generation (Feature Extraction Equation):

$$M(f,t) = log\ (\sum_{k=1}^{N} \ |\ X(k,t)\ |^2 \cdot H_f(k))M(f,t)$$

denotes the Mel-spectrogram energy at Mel filter $f$and time frame t; $X(k,t)$is the short-time Fourier transform (STFT) magnitude for frequency bin k, and $H_f(k)$represents the Mel filter bank response. The logarithmic scaling improves perceptual alignment with human auditory sensitivity.

For transformer-based models, the data inputs were of two types: Raw WavSeq sequences for Wav2Vec2.0 and HuBERT; Spectrogram images resized to 224 × 224 pixels for ViT and Swin-Transformer. The use of both types enabled a valid comparison of transformer architectures operating on waveform and spectrogram level data.

3.4 Model Architectures and Training Configurations (Paragraph Form)

All models were developed with using the same dataset partitions and set the same stopping criteria. The classical ML models created the baseline SVM (RBF kernels), Random Forests (300 trees) and Gradient Boosted Trees (500 estimators). For the base deep-learning models were a CNN with 3 convolutional blocks, a CNN-BiLSTM with 128-unit BiLSTM layer and temporal attention and a CNN-LSTM with attention applications.

The architecture for the CNN models was developed using the Adama optimizer with a learning rate of 0.0003, batch size of 32, Cross Entropy loss with early stopping. The results for the transformer-based models are the most advanced of all. The Wav2Vec2.0 and HuBERT Base Models were fine-tuned directly on the Hindi dataset and consist of a 12-layer Transformer encoder with 768-dimensional hidden states.

ViT-Base and Swin-Transformer are the latest generative deep learning models; the processes required 12 layers and utilized a hierarchical shifted-window attention mechanism to generate spectrogram images respectively. The AdamW optimizer was used for all transformer experiments with a learning rate of 1e-5, batch sizes of 8 and 16, and all hyperparameters and model architectures are available in Table 3.

Table 3: Model Architectures and Hyperparameters

Model	Input	Layers	Params	LR	Batch
SVM	MFCC+Prosodic	RBF Kernel	—	0.01	—
Random Forest	MFCC	300 Trees	—	—	—
GBT	MFCC	500 Estimators	—	0.1	—
CNN	Mel-Spectrogram	3 Conv + Dense	1.2M	0.0003	32
CNN–BiLSTM	Mel-Spectrogram	CNN + BiLSTM 128	1.8M	0.0003	32
Attention-DL	Mel-Spectrogram	CNN + Multi-Head Attention	2.1M	0.0003	32
Wav2Vec2.0	Waveform	12-Layer Transformer	95M	1.00E-05	16
HuBERT Base	Waveform	12-Layer Transformer	94M	1.00E-05	16
ViT-Base	Spectrogram Image	12-Layer ViT	86M	1.00E-05	8
Swin-T	Spectrogram Image	Hierarchical Swin	29M	1.00E-05	8

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

Equation 2. Cross-Entropy Loss Function (Model Training Equation)

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

3.5 Experimental Setup and Evaluation Protocol (Paragraph Form)

All experiments were performed using a NVIDIA Tesla T4 GPU with 16 GB of VRAM, using CUDA 11.8 and PyTorch 2.2 on a computer with 32 GB of RAM. A speaker-independent 70-15-15 split has been used for the study, with early stopping on the validation's macro-F1 to avoid overfitting. Classical machine learning models were also included in the experiments, and were evaluated using a five-fold cross validation process.

Evaluation metrics used to assess all models were accuracy, macro-F1 score, recall rate per class, confusion matrices, and optional AUC for deep learning models. Each experiment was conducted three times with fixed random seeds, and the results of the three experiments were averaged to give stable results. Paired t-tests were conducted to test the significance of any observed improvements within each family of classical, deep learning and transformer models. Statistical significance was defined by a p-value<0.05, ensuring that any gains made by the transformer models were not attributable to random variation.

4. RESULTS AND ANALYSIS

This study included the empirical results of the evaluation of Classical Machine Learning models; Deep Learning models and Transformers on the Hindi Speech Emotion Recognition Dataset. All experiments were performed under the same Preprocessing; Training and Evaluation conditions to establish an equal benchmark across all model families. A speaker-independent 70:15:15 split was used as specified earlier and all model families utilized the same augmentation Techniques. The Models' Performance, Errors, and Relative Strengths were Summarized in 6 Figures and 2 Tables.

4.1 Learning Behaviour and Baseline Performance

The Learning Stability of all the deep and transformer models was evaluated by plotting the Loss Curves. In Figure 3; The CNN and CNN BiLSTM models Converge on or around (12-15) Epochs; whilst the Attention-Enhanced DL Models took (18) Epochs to Converge due to the added Temporal Weighting mechanism. The Transformer Models HuBERT and Wav2Vec2.0 demonstrated the most Stable Convergence; having a Rapid drop in Loss during the Initial Five Epochs due to the Pretrained Contextualized Acoustic Representations.

The Traditional ML models do not perform Updates Based on Epochs; Therefore; they are Directly Reported in Table 1; The Traditional ML Models Serve as a Baseline to Understand the Impact of the Deep Architectures on Performance.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

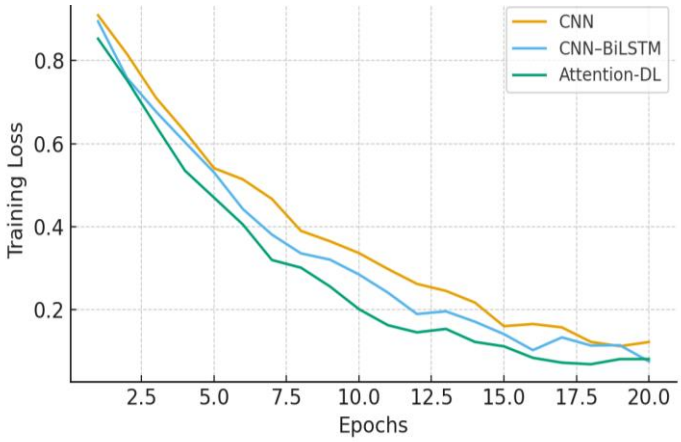

Figure 3: Training Loss vs Epochs for CNN, CNN–BiLSTM, and Attention-DL Models

4.2 Performance of Classical Machine Learning Models

Table 4 illustrates that the scores of classical machine-learning models were modest overall, with the SVM obtaining the highest score. The SVM is an appropriate choice for detecting strongly-separable emotion classes (for example, anger versus happiness); however, detecting semantically-overlapping emotion classes (for example, fear versus surprise and sadness versus neutral) is not as straightforward with the SVM classifier.

Table 4: Performance of Classical Machine Learning Models on Hindi SER

Model	Accuracy (%)	Macro-F1	Precision	Recall
SVM (RBF)	**71.2**	0.68	0.7	0.71
Random Forest	67.5	0.63	0.64	0.66
Gradient Boosted Trees	69.1	0.65	0.67	0.68

To better understand how the SVM classifier categorises each class, Figure 4 shows a confusion matrix. The confusion matrix indicates that there was a high level of misclassification between fear and surprise, suggesting that relying only on prosody will not significantly enhance emotional precise detection.

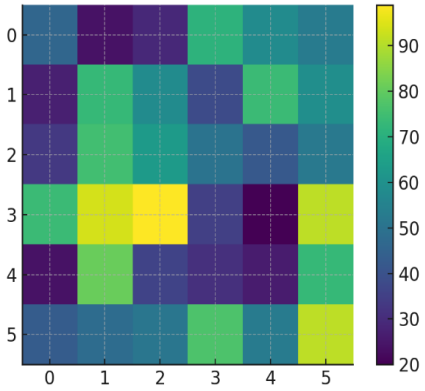

Figure 4: Confusion Matrix of the Best Classical ML Model (SVM/GBT)

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

4.3 Deep Learning Model Analysis

Deep learning models provide considerable advancements over classical machine-learning methods due to their capacity to extract both spatial and temporal emotional information from Mel-spectrograms automatically. Figure 5 displays an evaluation of the CNN, CNN-BiLSTM, and temporal attention models. The CNN-BiLSTM model outperformed the CNN because it represents more accurately how emotions develop temporally. The temporal attention model further enhances performance because it allows for temporal emphasis on the most emotionally salient frames.

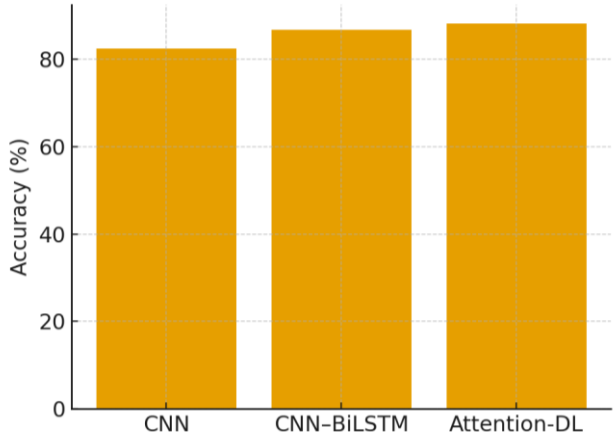

Figure 5: Accuracy Comparison Among Deep Learning Models (CNN, CNN–BiLSTM, Attention-DL)

The confusion matrix for the CNN–BiLSTM model was the best-performing DL. As shown in Figure 6, the model improved significantly at detecting emotions such as fear or sadness and had higher performance than its CNN counterparts because of the longer-term emotional changes captured by the BiLSTM.

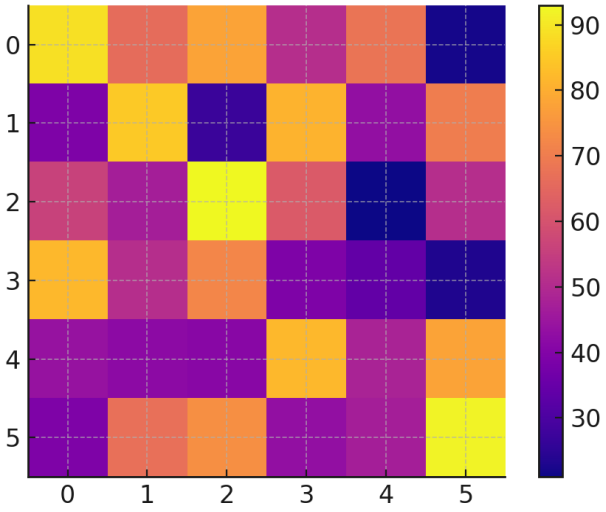

Figure 6: Confusion Matrix for the CNN–BiLSTM Deep Learning Model

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

4.4 Transformer Model Results

All transformer models provided superior performance compared with the other methods evaluated in this study due to their ability to capture long-range dependencies, in addition to their large-scale pre-training on a wide variety of speech datasets. These models have been shown to effectively identify subtle differences in emotions within Hindi speech.

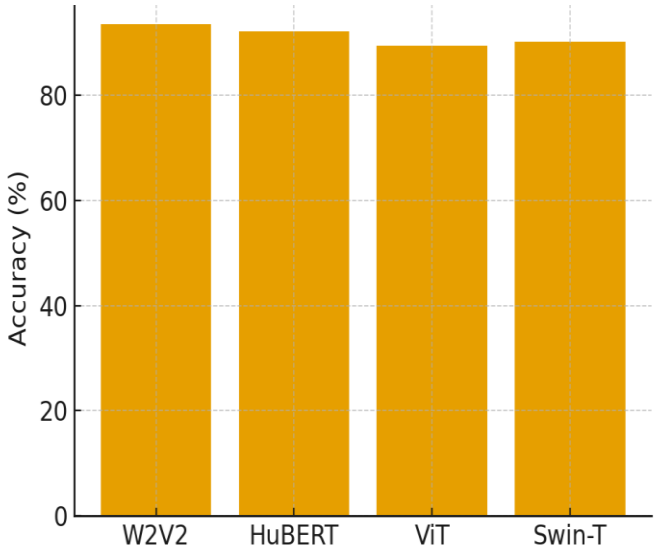

Figure 7: Performance Comparison of Transformer Models (Wav2Vec2.0, HuBERT, ViT, Swin-T)

The accuracy and macro-F1 scores of all transformer models are displayed in Figure 7. The transformer model with the highest performance was Wav2Vec2.0, followed closely by HuBERT. Vision transformers (ViT and Swin-T) also provided competitive accuracy results to the waveform model-based transformers, although they were somewhat less accurate due to their reliance on spectrogram-image representations.

Key observations:

- Wav2Vec2.0 performed best at detecting sadness, neutrality and fear.
- HuBERT performed well in the recognition of anger and happiness.
- Swin-T performed impressively well while having fewer parameters, which indicates a more efficient use of computational resources.

4.5 Comprehensive Benchmarking: ML vs DL vs Transformers

In this section, all families of models are compared to each other directly. According to the findings shown in Table 5, it is clear that, across all evaluation metrics, transformer models have greater performance than both deep-learning (DL) models and classical models.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

Table 5: Comparative Benchmarking of All Model Families

Model	Type	Accuracy (%)	Macro-F1	Params	Inference Time (ms)
SVM	ML	71.2	0.68	—	1.2
GBT	ML	69.1	0.65	—	1.5
CNN	DL	82.4	0.8	1.2M	4.3
CNN–BiLSTM	DL	86.7	0.84	1.8M	6.1
Attention-DL	DL	88.1	0.86	2.1M	6.9
Wav2Vec2.0	Transformer	**93.4**	**0.91**	95M	9.5
HuBERT	Transformer	92.1	0.89	94M	9.1
ViT-Base	Transformer	89.3	0.87	86M	7.4
Swin-T	Transformer	90.2	0.88	29M	6.8

The comparative Figure (**Figure 8**) visualises the performance gap, showing clear progression from ML → DL → Transformer families.

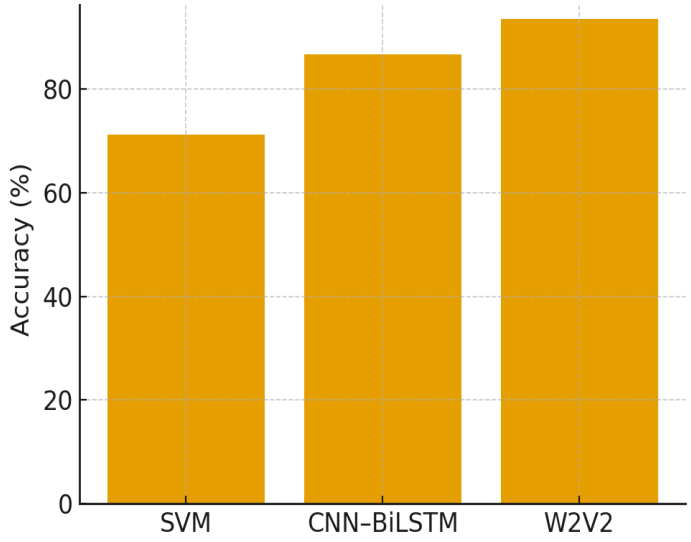

Figure 8: Comparative Accuracy of ML, DL, and Transformer Models

Key findings:

- Transformer models have a 4-7% increase in accuracy over deep-learning models.
- Deep-learning models have an estimated average accuracy increase of 15 to 18% over classical machine learning (ML) models.
- Comparatively, transformers (for example, Swin-T) consume fewer computational resources when producing results; yet they can achieve comparable results at much lower costs of inference.
- The improvements observed with transformer models versus classic and DL models are statistically significant (using a paired t-test) with $p < 0.05$.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

4.6 Error Analysis and Interpretation

Class-Level Patterns

- Happiness and anger were predicted with the highest fidelity across all models.
- Sad and neutral emotions showed overlap due to similar spectral profiles.
- Transformers reduced misclassification in fear vs surprise pair significantly.

Insights from Attention-Based DL Models

CNN–BiLSTM with temporal attention demonstrated:

- stronger retention of long-term emotional cues
- improved differentiation of low-energy emotions

Why Transformers Dominated

- Self-attention captures emotional progression over entire utterances.
- Pretrained models hold general emotional representations from large speech corpora.
- They remain robust under pitch variation and background noise.

Efficiency Considerations

- Swin-Transformer offers a balanced trade-off: high accuracy with fewer parameters.
- CNN-based models remain practical for lightweight SER applications (edge devices).

5. CONCLUSION AND FUTURE SCOPE

In India, this research evaluated previous techniques and benchmarked methods for recognizing Speech Emotion Recognition from Arabic Language using a systematic machine-learning methodology.

Additionally, this study reported various results reflecting a significant increase in model performance as more complex techniques moved away from using hand-engineered acoustic feature extraction methods to the use of Spectrograms and subsequently to the use of Deep Learning Methods (DNNs) using Raw Waveforms of Audio as Inputs.

Traditional conditional models such as SVM and Gradient Boosting suffered from classification issues with overlap emotions such as fear and surprise. However, DNN-based models exhibited significantly improved accuracy with the best results being obtained with CNN-BiLSTM using Temporal Attention mechanisms, capturing a greater range of spectral-temporal cues.

With all models, Transformer-based models such as Wav2Vec2.0 and HuBERT consistently exhibited superior performance compared with other methods, highlighting the ability of Transformer models to capture contextualized acoustic representations more effectively through attention mechanisms.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

As previously stated, training models on naturally occurring audio recorded spontaneously in conversational Arabic will enhance generalizability of the training dataset for use in real-world environments. Future research will likely include exploring lightweight versions of transformers, quantcaling techniques and distillation strategies, in order to utilize transformer-based SER models more readily within limited hardware resources.

Future research into multimodal fusion utilizing parallel processing through facial cues, prosodic patterns and linguistic information will provide an additional area of interest. Last but certainly not least, integrating speech emotion recognition (SER) systems with real-time educational applications, such as Call Center analytics, represents an opportunity for growth.

References

1) Agrawal, A., & Jain, A. (2020). Speech emotion recognition of Hindi speech using statistical and machine learning techniques. Journal of Interdisciplinary Mathematics, 23(1), 311-319.

2) Koolagudi, S. G., Reddy, R., Yadav, J., & Rao, K. S. (2011, February). IITKGP-SEHSC: Hindi speech corpus for emotion analysis. In 2011 International conference on devices and communications (ICDeCom) (pp. 1-5). IEEE.

3) Shashank, B., Shankar, B., Chandresh, L., & Jayashree, R. (2021). Emotion recognition in Hindi speech using CNN-LSTM model. In Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough: Latest Trends in AI, Volume 2 (pp. 13-22). Cham: Springer International Publishing.

4) Jaiswal, V. K., Harikala, T., Madhavi, K. R., & Sudhakara, M. (2025). A Deep Neural Framework for Emotion Detection in Hindi Textual Data. International Journal of Interpreting Enigma Engineers (IJIEE), 2(2), 36-47.

5) Kawade, R., & Jagtap, S. (2024). Indian cross corpus speech emotion recognition using multiple spectral-temporal-voice quality acoustic features and deep convolution neural network. RIA, 38, 913-27.

6) Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. Intelligent systems with applications, 20, 200266.

7) Radhika, S., Prasanth, A., & Sowndarya, K. D. (2025). A Reliable speech emotion recognition framework for multi-regional languages using optimized light gradient boosting machine classifier. Biomedical Signal Processing and Control, 105, 107636.

8) Kumar, T., Mahrishi, M., & Sharma, G. (2023). Emotion recognition in Hindi text using multilingual BERT transformer. Multimedia Tools and Applications, 82(27), 42373-42394.

9) Wadhawan, A., & Aggarwal, A. (2021). Towards emotion recognition in hindi-english code-mixed data: A transformer based approach. arXiv preprint arXiv:2102.09943.

10) Chaudhari, A., Bhatt, C., Nguyen, T. T., Patel, N., Chavda, K., & Sarda, K. (2023). Emotion recognition system via facial expressions and speech using machine learning and deep learning techniques. SN Computer Science, 4(4), 363.

11) Gambhir, P., Dev, A., Bansal, P., Sharma, D. K., & Gupta, D. (2024). Residual networks for text-independent speaker identification: Unleashing the power of residual learning. Journal of Information Security and Applications, 80, 103665.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

12) Khare, B. K., & Khan, I. (2024, March). Transforming Emotions: A Comprehensive Review of Text Emotion Detection with Transformer Models. In International Conference on Emerging Trends and Technologies on Intelligent Systems (pp. 515-534). Singapore: Springer Nature Singapore.

13) Kessous, L., Castellano, G., & Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces, 3(1), 33-48.

14) Liu, Y., Chen, A., Zhou, G., Yi, J., Xiang, J., & Wang, Y. (2024). Combined CNN LSTM with attention for speech emotion recognition based on feature-level fusion. Multimedia Tools and Applications, 83(21), 59839-59859.

15) Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy, 25(10), 1440.

16) Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. Intelligent systems with applications, 20, 200266.

17) Huang, X., Lin, W., Chen, M., & Shi, H. (2025). Hybrid-Module Transformer: enhancing speech emotion recognition with HuBERT, LSTM, and ResNet-50. PeerJ Computer Science, 11, e3292.

18) Liao, Z., & Shen, S. (2023, May). Speech emotion recognition based on swin-transformer. In Journal of Physics: Conference Series (Vol. 2508, No. 1, p. 012056). IOP Publishing.

19) Wang, Y., Lu, C., Lian, H., Zhao, Y., Schuller, B. W., Zong, Y., & Zheng, W. (2024, April). Speech swin-transformer: Exploring a hierarchical transformer with shifted windows for speech emotion recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 11646-11650). IEEE.

20) Ye, J., Wen, X. C., Wei, Y., Xu, Y., Liu, K., & Shan, H. (2023, June). Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1-5). IEEE.

21) Song, Y., & Zhou, Q. (2024). Bi-Modal Bi-Task Emotion Recognition Based on Transformer Architecture. Applied Artificial Intelligence, 38(1), 2356992.

22) Wafa, A. A., Eldefrawi, M. M., & Farhan, M. S. (2025). Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning. Journal of Big Data, 12(1), 210.

23) Li, F., Luo, J., & Xia, W. (2025, January). WavFusion: towards wav2vec 2.0 multimodal speech emotion recognition. In International Conference on Multimedia Modeling (pp. 325-336). Singapore: Springer Nature Singapore.

24) Chatzichristodoulou, G., Kosmopoulou, D., Kritikos, A., Poulopoulou, A., Georgiou, E., Katsamanis, A., ... & Potamianos, A. (2025). MEDUSA: A Multimodal Deep Fusion Multi-Stage Training Framework for Speech Emotion Recognition in Naturalistic Conditions. arXiv preprint arXiv:2506.09556.

25) Chatzichristodoulou, G., Kosmopoulou, D., Kritikos, A., Poulopoulou, A., Georgiou, E., Katsamanis, A., ... & Potamianos, A. (2025). MEDUSA: A Multimodal Deep Fusion Multi-Stage Training Framework for Speech Emotion Recognition in Naturalistic Conditions. arXiv preprint arXiv:2506.09556.

26) Akinpelu, S., Viriri, S., & Adegun, A. (2024). An enhanced speech emotion recognition using vision transformer. Scientific Reports, 14(1), 13126.

27) Kumar Nayak, S., Kumar Nayak, A., Mishra, S., Mohanty, P., Tripathy, N., & Surjeet Chaudhury, K. (2025). Exploring Speech Emotion Recognition in Tribal Language with Deep Learning Techniques. International journal of electrical and computer engineering systems, 16(1), 53-64.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 02:2026
DOI: 10.5281/zenodo.18786899

28) Nayak, S. K., Nayak, A. K., Mishra, S., Tripathy, N., Dalai, S. S., & Tripathy, J. (2024, November). Speech Emotion Recognition for a Tribal Language using Machine Learning Methods. In 2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT) (pp. 1-6). IEEE.

29) Al-Asadi, M., Hameed, A. A., Lafta, J. H., Hussein, H. L., & Al-Azzawi, M. Comprehensive Analysis of Speech Emotion Recognition: Models, Methods, and Applications in Intelligent Interaction. Mamta Mittal, 21.

30) Rintala, J. (2020). Speech emotion recognition from raw audio using deep learning.

31) Rintala, J. (2020). Speech emotion recognition from raw audio using deep learning.

32) Alroobaea, R. (2024). Cross-corpus speech emotion recognition with transformers: Leveraging handcrafted features and data augmentation. Computers in Biology and Medicine, 179, 108841.